FORENSIC EVALUATION OF THE EVIDENCE USING AUTOMATIC SPEAKER RECOGNITION SYSTEMS

–TESIS DOCTORAL–

EVALUACIÓN DE LA EVIDENCIA FORENSE UTILIZANDO SISTEMAS AUTOMÁTICOS DE RECONOCIMIENTO DE LOCUTOR

Author: Daniel Ramos Castro
(Ingeniero de Telecomunicación, Universidad Politécnica de Madrid)

A Thesis submitted for the degree of:

Doctor of Philosophy

Madrid, November 2007
Colophon

This book was typeset by the author using \LaTeXe. The main body of the text was set using a 11-points Computer Modern Roman font. All graphics and images were included formatted as Encapsuled Postscript (TM Adobe Systems Incorporated). The final postscript output was converted to Portable Document Format (PDF) and printed.

Copyright © 2007 by Daniel Ramos Castro. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author. Universidad Autonoma de Madrid has several rights in order to reproduce and distribute electronically this document.
The research described in this Thesis was carried out within the Biometric Recognition Group – ATVS at the Dept. of Ingeniería Audiovisual y Comunicaciones, Escuela Universitaria de Ingeniería Técnica de Telecomunicación, Universidad Politécnica de Madrid (from 2003 to 2005); and at the Dept. of Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2005 to 2007). The project was partially funded by a PhD scholarship from Comunidad de Madrid and Fondo Social Europeo.
The author was awarded with a PhD scholarship from Consejería de Educación de la Comunidad de Madrid and Fondo Social Europeo between 2003 and 2006 which partially supported the research summarized in this Dissertation.


Abstract

This Thesis is focused on the use of automatic speaker recognition systems for forensic identification, in what is called forensic automatic speaker recognition. More generally, forensic identification aims at individualization, defined as the certainty of distinguishing an object or person from any other in a given population. This objective is followed by the analysis of the forensic evidence, understood as the comparison between two samples of material, such as glass, blood, speech, etc. An automatic speaker recognition system can be used in order to perform such comparison between some recovered speech material of questioned origin (e.g., an incriminating wire-tapping) and some control speech material coming from a suspect (e.g., recordings acquired in police facilities).

However, the evaluation of such evidence is not a trivial issue at all. In fact, the debate about the presentation of forensic evidence in a court of law is currently a hot topic in many scientific and legal fora. The American Daubert rules for the admissibility of the scientific evidence in trials and the evidence of critical errors in positive identification reports for disciplines assumed as error-free have fostered the discussion. From this debate, DNA profiling arises as a model for a scientifically defensible approach in forensic identification, as it meets the most stringent Court admissibility requirements demanding scientific evaluation of the evidence, and testability of procedures. In this Thesis we take into account such requirements in order to adapt forensic automatic speaker recognition to what has been dubbed the coming paradigm shift in forensic identification science.

We begin by reviewing related works in the literature concerning automatic speaker recognition and forensic evaluation of the evidence. Then, the experimental framework to be used in this Thesis is described in detail. The widely accepted Speaker Recognition Evaluations (SRE) conducted by the American National Institute of Standards and Technology (NIST) are adopted as the experimental set-up for this Thesis. The databases used for such protocols constitute challenging corpora presenting many different variability factors, simulating the typical conditions of lawful recordings in telephonic networks.

As a contribution in this Thesis, a hierarchical methodology for forensic automatic speaker recognition is proposed. This methodology constitutes a powerful tool for practitioners, as it allows transparent and testable forensic identification using the typical score-based automatic speaker recognition systems. We then identify the main factors affecting the methodology proposed in this Thesis. First the elements of the coming paradigm shift are analyzed. Then, the common procedures accepted in automatic forensic speaker recognition are also identified. Taking into account all factors, we define the hierarchical methodology, consisting of three different levels of abstraction, namely the discrimination level, the presentation level and the forensic level.

The Dissertation then focuses on the description of the levels which compose the proposed
hierarchical methodology. First, the discrimination level is addressed. The aim at this level is to yield a discriminating score, as a way of distinguishing whether the speech coming from the suspect and the questioned recording come from the same source or not. Since discrimination has been the aim of automatic speaker recognition in the last decades, we give a definition of the performance of the score derived from the literature in the field. Moreover, we overview and experimentally compare several widely used techniques found in the literature in order to improve the discriminating power of a score set, namely score normalization, session variability compensation and fusion of systems. A novel score normalization technique, namely KL-T-Norm, is presented as a contribution. We experimentally demonstrate that KL-T-Norm increases the discriminating power of other popular score normalization techniques such as T-Norm, as well as it improves its computational efficiency.

Next, the presentation level is introduced. The aim at this level is transforming the input score into a likelihood ratio (LR) as a measure of the weight of the evidence, with a meaning of degree of support of the evidence to any of the hypotheses present in the case. This methodology, popularized by DNA profiling, is probabilistic, data-driven and allows to include in a logical way the weight of the evidence into the inferential process in a forensic case. A definition of the accuracy of the evidence evaluation process is then given, introducing the important concept of calibration. Then, a novel assessment methodology based on information theory is reported, where the accuracy of the LR values is expressed in the form of information-theoretical magnitudes, namely empirical cross-entropy (ECE).

Also in the presentation level, a comparative study of different LR computation techniques is presented. Among them, we propose a novel method of generative suspect-adapted LR computation. The study shows that the proposed technique improves the discrimination and the calibration of the input scores, by means of the exploitation of the specificities of a given suspect. The proposed technique is also robust to scarcity in the control speech material, a problem which is often found in forensic casework. The presentation level is concluded with an alternative configuration of the proposed methodology in order to consider non-score-based LR computation techniques, common in other forensic areas and recently proposed for automatic speaker recognition.

Finally, the last level in the hierarchy is described, namely the forensic level. The aim at this level is considering the court demands and the requirements of the coming paradigm shift in forensic science in order to properly report the weight of the evidence and its accuracy. Two experimental examples illustrate the reporting and presentation of the results from evidence evaluation by means of the proposed information-theoretical assessment methodology. One of these examples has been built making use of the database and systems employed by the Spanish Guardia Civil in real forensic casework. The chapter ends with the demonstration of the adequacy of the proposed methodology for other forensic disciplines, by means of an experimental example of LR-based evidence evaluation using glass and paint analysis.
A mis padres y a mi hermano.

Our aim as scientists is objective truth, more truth, more interesting truth, more intelligible truth. We cannot reasonably aim at certainty. Once we realize that human knowledge is fallible, we realize also that we can never be completely certain that we have not made a mistake.

Acknowledgements

This PhD Thesis summarizes the work carried out during my Ph.D. studies with the ATVS - Biometric Recognition Group since 2003. This research group was established in 1994 at the Dept. of Ingeniería Audiovisual y Comunicaciones (DIAC) of the Universidad Politécnica de Madrid (UPM) and since 2004 is affiliated to the Dept. of Ingeniería Informática of the Universidad Autónoma de Madrid (UAM). The work presented in this Thesis has been conducted at both institutions. The financial support for the first year of the Ph.D. studies came from a research grant with Ministerio del Interior (Dirección General de la Guardia Civil, DGGC). Subsequent three years have been financially supported by a Ph.D. scholarship from Comunidad de Madrid and Fondo Social Europeo, and various Spanish and European projects. During the last year of my studies, I have taken benefit from a Teaching Assistant position at Escuela Politécnica Superior, Universidad Autónoma de Madrid.

Foremost, I would like to thank my advisor Prof. Joaquín González-Rodríguez for his guidance and support over the past five years. He is the responsible of my first steps in the research world, and the confidence he has always put in me has definitely fostered my motivation. During these years I have benefited from his guidance, vision, discipline and intelligent effort, which have shaped my working attitudes. In the framework of the ATVS research group, I have received support from Prof. Javier Ortega-García, feeling fortunate to learn from his courage and self-mastery. I have been also advised by Prof. Luis A. Hernández during the obtention of my DEA (Diploma de Estudios Avanzados), having the opportunity to experience his exemplifying and encouraging philosophy of research work. I also want to thank Prof. José Colás, for his work on reviewing this Thesis, and for his wise comments and suggestions; as well as Juana Calle, for her patience and her valuable help.

During the journey of my Ph.D. degree I have been fortunate to meet many excellent professionals and colleagues. From DIAC at the UPM, I had the pleasure of meeting Dr. Danilo Simón-Zorita, Dr. José Luis Sánchez-Bote, Dr. Antonio Pedrero and Dr. Luis I. Ortiz, having illustrating discussions and receiving valuable advice from them. After moving to the UAM and becoming a Teaching Assistant, I was lucky to learn from colleagues like Dr. Ángel de Castro, Dr. Nicolás Morales, Dr. Kurosh Khan-Afshar and Dr. Pablo Haya, as well as Profs. Javier Garrido, Susana Holgado, Antonio Aguilar and Juan-Alberto Sigüenza.

The opportunity of visiting foreign institutions during my Ph.D. studies was a real luck. I especially remember my first research stay at the University of Lausanne. I had the fortune to meet Prof. Christophe Champod there, whose outstanding work in the area of forensic identification was the inspiration of this Ph.D. Thesis. His extraordinary care as a host during my visit and his intelligence and constant support during those three months was essential for the development of this work. Also, he was partly responsible of my happiness during the stay, with his kind relationship and help, together with his colleagues at the University, Anne
Marville and Profs. Pierre Margot, Franco Taroni, Olivier Ribaux and Andrzej Drygajlo. My second stay was in the University of Edinburgh, with Prof. Colin G. G. Aitken. During those three months I received an enormously motivating support and confidence from Prof. Aitken, and I could benefit from his mastery in the field of statistics for evidence evaluation, which led to some contributions contained in this Thesis. I had also the opportunity to meet Dr. David Lucy and Dr. Burkhard Schaffer, whose kindness and wisdom helped me in daily work. Finally I stayed three months in Spescom Datavoice and the University of Stellenbosch, South Africa, where I had the luck to meet Niko Brümmer and Prof. Johan du Preez. Some of the most important contributions in this Thesis are based on their work on the evaluation of speaker detection performance, which reflect their open scientific mind and their outstanding capabilities for solving challenging problems in the research community. I also thank Viv Crone and Sakkie Nel for their help and interest on my work during the stay.

From such research stays, there is a huge number of colleagues and friends who have awarded my life with their presence. With apologies to all those wonderful people not mentioned here, I have to especially thank Line Gueissaz, Dr. Olivier Delemont, Dr. Pierre Esseiva, Romain Voisard, Dr. Plamen Prodanov, Dr. Jonas Richiardi, Dr. Krzysztof Krzyzczuk, The Keans (Norie and Wyanette), Arturo Belmonte and Sanet Stander and her wonderful family. All of them, among many others, have made possible the difficult task for a Spaniard to stop missing his country at least for a while. I have to especially thank die Brümmer for their friendship and for hosting me as one more member in their family. I lived with them some of the sweetest moments I ever had. Piet, Niklas, Jan, Madrie and Niko: dankie!

I would like also to acknowledge a number of researchers which have helped to shape this PhD Thesis with their interesting discussions, and who have motivated me with their support and inspiring work. These researchers include Dr. Phil Rose, Dr. Grzegorz Zadora, Dr. David van Leeuwen, Prof. Hermann Künzell, Dr. Jos Bouten and Dr. Didier Meuwly. I would also like to thank people from Guardia Civil, especially Major José-Juan Lucena, for their constant support and confidence.

It is mandatory for me to thank all the work mates at ATVS who have shared with me tons of time, stress, achievements, friendship, bad moods and laughs, and who daily enrich my professional and personal life. I have to wholeheartedly thank (in non-perfect order of appearance): Daniel G. Romero, Marta G. Gomar, Dr. Julián Fírreiz, Marta Sánchez, Ignacio López, Javier Franco, Alberto Montero, Fernando Alonso, Jaime López, Javier G. Torcelly, Prof. Doroteo T. Toledano, Carlos Bousoño, Javier González, Javier Simón, Lucas Pérez, Javier Galbally, Manuel Freire, Víctor M. González, Danilo Spada, Cristina Esteve, Alejandro Abejón and Ismael Mateos, among many others who have recently joined this exciting research journey at ATVS. Special thanks are given to Prof. Doroteo T. Toledano, Manuel Freire and Dr. Julián Fírreiz for their valuable work on the review process of this Thesis, and to Javier and Santiago González for sharing pains with me in the last stages.

Es imposible para mí resumir en unas simples líneas todo lo que siento por vosotros, todas
esas personas que giráis alrededor de mi vida fuera de lo profesional, y sin las cuales todo este trabajo no tendría sentido alguno. La felicidad que me llena cuando me doy cuenta de la titánica tarea que es pasaros lista es sólo comparable al amor que os profeso por todo lo que habéis hecho y hacéis en mi vida. Citaros a todos sería improbo más posible. Expresar en papel lo que siento por vosotros es para mí irrealizable. Permitidme entonces que os lo agradezca así, a todos juntos y a la vez. Porque sería irremediable e imperdonable olvidarme de alguien. Porque sabéis quiénes sois y que yo sé quiénes sois. Porque no sabría cómo ordenaros por orden de importancia.

Hay sin embargo tres personas a las que cito explícitamente, y a las que va dedicada esta Tesis. Son mis padres y mi hermano. Todo lo que soy os lo debo a vosotros.

Daniel Ramos Castro

Madrid, November 2007
Contents

Abstract vii

Acknowledgements xi

List of Figures xviii

List of Tables xxv

1. Introduction 1
   1.1. Forensic identification science in the 21st century 4
       1.1.1. Classical forensic identification 6
       1.1.2. The need of a paradigm shift 7
       1.1.3. Emulating DNA in forensic automatic speaker recognition 9
   1.2. Motivation of this PhD Thesis 9
   1.3. The Thesis 11
   1.4. Outline of the Dissertation 12
   1.5. Research contributions 15

2. Related works 19
   2.1. Automatic speaker recognition 19
       2.1.1. Identity information in the speech signal 19
       2.1.2. Applications 21
       2.1.3. Technology 21
       2.1.4. Score-based architecture 22
       2.1.5. Feature extraction and tokenization 23
       2.1.6. Text-independent speaker recognition 25
   2.2. LR–based analysis of the evidence 30
       2.2.1. Historical review 30
       2.2.2. The LR approach for evidence evaluation 31
       2.2.3. LR computation in forensic science 31
   2.3. LR computation using automatic speaker recognition systems 33
       2.3.1. First approaches to LR–based forensic automatic speaker recognition 34
2.3.2. Robustness in LR–based forensic automatic speaker recognition
2.3.3. Discriminative LR computation in forensic automatic speaker recognition
2.4. Chapter summary and conclusions

3. Experimental framework
3.1. Evaluation methodology
3.1.1. Statistical significance of experimental results
3.2. Experimental framework
3.2.1. Databases
3.2.2. Experimental protocols
3.2.3. Baseline automatic speaker recognition systems
3.3. Chapter summary and conclusions

4. A hierarchical methodology for forensic automatic speaker recognition
4.1. The coming paradigm shift in forensic identification science
4.2. Forensic automatic speaker recognition in the steps of DNA
4.3. Common procedures in automatic speaker recognition
4.4. Hierarchical methodology for forensic automatic speaker recognition
4.4.1. Hierarchy of abstraction levels
4.5. Chapter summary and conclusions

5. The discrimination level: score-based automatic speaker recognition
5.1. Discriminating power
5.1.1. Discriminating power of a set of scores
5.1.2. Evaluating the discriminating power
5.2. Improving discriminating power in automatic speaker recognition
5.2.1. Score normalization
5.2.2. Contribution: KL-T-Norm: speaker- and test- dependent fast score normalization
5.2.3. Session variability compensation
5.2.4. Fusion
5.3. Chapter summary and conclusions

6. The presentation level: LR computation
6.1. Evaluation of the evidence using LR values
6.1.1. The meaning of the LR
6.2. Bayesian decision theory
6.3. The accuracy of the LR
6.3.1. Calibration and refinement of LR values
6.3.2. Assessing LR accuracy in forensic automatic speaker recognition
6.3.3. Examples of the effects of calibration
6.4. Contribution: information-theoretical evaluation of the accuracy of \( LR \) values .......................... 106
   6.4.1. Information-theoretical evaluation .......................................................... 107
   6.4.2. \( ECE \) and information loss ............................................................... 111
   6.4.3. The \( ECE \) Plot .............................................................................. 112
   6.4.4. Comparison to other performance measures ............................................. 115
   6.4.5. Experimental example ...................................................................... 115
6.5. Transforming scores into accurate \( LR \) values ......................................................... 116
   6.5.1. Pure score calibration techniques ............................................................. 118
   6.5.2. Hybrid score discrimination-calibration .................................................... 120
   6.5.3. Contribution: generative suspect-adapted \( LR \) computation .................. 122
   6.5.4. Experimental comparison of \( LR \) computation methods ......................... 128
6.6. \( LR \) computation for non-score-based systems .................................................... 130
6.7. Chapter summary and conclusions .................................................................. 130

7. The forensic level: forensic reporting and presentation in court ......................... 133
   7.1. Reporting to the court with forensic automatic speaker recognition systems .... 135
      7.1.1. The \textit{coming paradigm shift} and forensic automatic speaker recognition reporting .................................................. 135
      7.1.2. Speech variability and forensic automatic speaker recognition reporting . 136
   7.2. Presenting results in court using forensic automatic speaker recognition systems 137
      7.2.1. Experimental example: the use of the ATVS forensic speaker recognition system in a simulated case .................................................. 137
      7.2.2. Experimental example: simulated case with real forensic data and systems 140
   7.3. Generalization to other forensic disciplines ..................................................... 143
      7.3.1. The \( LR \) meaning and accuracy across forensic disciplines ................. 143
      7.3.2. Experimental example: glass and paint evidence .................................. 144
      7.3.3. Results ...................................................................................... 145
   7.4. Chapter summary and conclusions ................................................................ 148

8. Conclusions and future work .............................................................................. 149
   8.1. Conclusions ...................................................................................... 149
      8.1.1. Main results and contributions ............................................................. 151
   8.2. Future Work .................................................................................... 152
List of Figures

1.1. Information present in a case in order to answer the relevant question for the fact finder, which is usually “Do the control and recovered material come from the same source?”. ................................................................. 2

1.2. An example of error in a positive identification. On the left, latent fingerprint believed to belong to a terrorist involved in the train bombings in 11 March 2004 in Madrid, Spain. On the right, a fingerprint resulting from a database search and belonging to Brandon Mayfield from Portland, Oregon. On the basis of these two prints the FBI fingerprint examiners erroneously identified Mayfield as the source of the latent print. 15 characteristic points (minutiae) are marked in both images. Source: DoJ, 2006. ................................................................. 3

1.3. Transfer evidence. ................................................................. 5

1.4. Dependence among Dissertation chapters. ................................. 14

2.1. Identity levels in the speech signal. Source: Reynolds et al. (2003). ........ 20

2.2. Score-based speaker recognition system: (a) enrollment, (b) score computation. . 23

2.3. General scheme for a speaker verification system. ............................. 23

2.4. Short-term feature extraction. .................................................. 24

2.5. Example of pdf from a 4− mixture, 2− dimensional GMM. ..................... 26

2.6. Score computation in phonotactic speaker recognition. .......................... 28

2.7. Prosodic tokenization. ............................................................ 28

2.8. Bayesian inferential framework in LR-based evidence analysis. .................... 32

2.9. LR computation following the methodology proposed by Meuwly [2001] for a GMM-based automatic speaker recognition system. The recovered speech material is referred to as the trace here. Source: Gonzalez-Rodriguez et al. [2006]. . . . 35
2.10. Generative LR computation. In (a), the within- and between-source distributions \((p(e|θ_p)\) and \(p(e|θ_d)\) respectively) are obtained from the automatic speaker recognition system scores according to Figure 2.9. The LR will be the quotient of such distributions at a given value \(E\) of the evidence \(e\). \(E\) is the value of the similarity score given by the system when comparing the control and recovered materials. In (b), the score to \(\log(LR)\) mapping function is shown. As the variances of the Gaussian distributions in (a) are equal, the mapping in (b) is linear, but it will be quadratic in general for Gaussian pdfs with different variances.

2.11. Variability of population scores depending on the transmission channel of the speech from the population. Each plot shows the Gaussian fit of the distributions of scores obtained with a recovered speech sample and two populations: a landline telephone population (solid curve) and a cellular telephone population (dashed curve). Scores obtained with the ATVS-UAM GMM system in NIST SRE 2006. Population data extracted from NIST SRE 2004 database.

2.12. Variability of within-source distribution due to control speech data scarcity. Each plot shows the Gaussian fit of the within-source pdfs of 3 scores obtained with a given amount of control speech material from a suspect. Solid and dashed within-source pdfs have been obtained from the same set of 3 scores except one of them, which is different. Scores obtained with the ATVS-UAM GMM system in NIST SRE 2006.

2.13. Score-based LR computation. First, a score is computed. Second, a LR is obtained from the score, the control recordings (C database) and the speech from the relevant population (P database). This scheme is common for generative and discriminative LR computation techniques.

4.1. Factors present in 85 wrongful convictions, based on the case analysis data from the Innocence Project. The two highlighted bars indicate causes related to forensic science. Source: Saks and Koehler [2005].

4.2. Interrelated needs identified for forensic disciplines to conform to the coming paradigm shift in forensic identification.

4.3. Identified lines to follow in the proposed methodology in order to adapt forensic automatic speaker recognition to the coming paradigm shift in forensic science.

4.4. Elements considered by the proposed methodology for forensic automatic speaker recognition.

4.5. Abstraction levels in the proposed methodology for forensic automatic speaker recognition.

5.1. The discrimination level in the proposed hierarchical methodology for forensic automatic speaker recognition.

5.2. FA and FR rates curves for two sets of scores presenting perfect (a) and non-perfect (b) discriminating power, with respect to the decision threshold \(τ\). 
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Example of detection performance with ROC (left) and DET curves (right).</td>
<td>65</td>
</tr>
<tr>
<td>5.4</td>
<td>Example of FA/FR rate plots of two score sets having the same discriminating power. Score set represented in (b) has been obtained applying an affine transformation (shifting and scaling) to the scores represented in (a). Two sample thresholds (vertical lines) have been plot in (a), and their corresponding thresholds are represented in (b).</td>
<td>66</td>
</tr>
<tr>
<td>5.5</td>
<td>Test-normalization technique (T-Norm).</td>
<td>70</td>
</tr>
<tr>
<td>5.6</td>
<td>KL-T-Norm.</td>
<td>73</td>
</tr>
<tr>
<td>5.7</td>
<td>EER performance in the development set. The number of models in the cohort, $K$, is represented in the horizontal axis for the GMM (left) and SVM-GLDS (right) systems and for 1c-1c (up) and 8c-1c (down) conditions.</td>
<td>74</td>
</tr>
<tr>
<td>5.8</td>
<td>minDCF performance (as defined by NIST) in the development set. The number of models in the cohort, $K$, is represented in the horizontal axis for the GMM (left) and SVM-GLDS (right) systems and for 1c-1c (up) and 8c-1c (down) conditions.</td>
<td>74</td>
</tr>
<tr>
<td>5.9</td>
<td>KL-T-Norm in blind NIST 2005 SRE. (a) GMM system and (b) SVM-GLDS system.</td>
<td>76</td>
</tr>
<tr>
<td>5.10</td>
<td>Effects of NAP channel compensation technique in short-time spectral systems, namely (a) a GMM-SVM-Supervectors system with feature mapping and (b) a GMM system with feature warping (for the latter, NAP is applied only to trained GMM models, not to test segments).</td>
<td>80</td>
</tr>
<tr>
<td>5.11</td>
<td>Effects of the fusion of ATVS GMM and SVM-GLDS systems for the 1conv4w-1conv4w condition of NIST SRE 2005 (a) and NIST SRE 2006 (b).</td>
<td>82</td>
</tr>
<tr>
<td>5.12</td>
<td>Effects of the fusion of ATVS-UAM spectral and higher-level systems for the 8conv4w-1conv4w condition of NIST SRE 2005 (a) and NIST SRE 2006 (b).</td>
<td>82</td>
</tr>
<tr>
<td>6.1</td>
<td>The presentation level in the proposed hierarchical methodology for forensic automatic speaker recognition.</td>
<td>86</td>
</tr>
<tr>
<td>6.2</td>
<td>Odds can be viewed as stones in a balance.</td>
<td>89</td>
</tr>
<tr>
<td>6.3</td>
<td>Three different cases illustrating posterior odds inference from the prior odds (province of the fact finder) and the evidence analysis (province of the forensic scientist).</td>
<td>90</td>
</tr>
<tr>
<td>6.4</td>
<td>Definition of the decision regions ($R_p$ and $R_d$) in Bayesian decision theory. (a) shows the value of the likelihoods of $e$ conditioned to $\theta_p$ and $\theta_d$ respectively. The boundary $B$ of the decision regions is placed according to the value of the likelihoods and Bayes decision rule (Equation 6.8), being $R_p$ the region of the $e$ space where $\theta_p$ is decided, and vice-versa. (b) shows the relationship between the decision boundary in the $e$ space and the logarithm of the $LR$. $\theta_p$ will be decided if $LR &gt; \frac{P(\theta_d)}{P(\theta_p)}$ and vice-versa. In this example, $(P(\theta_p) = P(\theta_d))$ and therefore the logarithm of the decision threshold is zero.</td>
<td>93</td>
</tr>
</tbody>
</table>
6.5. Definition of the decision regions ($R_p$ and $R_d$) in Bayesian decision theory when Bayes’ threshold $\tau_B$ (priors and costs) change. In (a) we show the likelihoods and decision regions in the $e$ space. If the decision boundaries are chosen according to Bayes decision rule (Equation 6.13) then an increase in $P(\theta_p)$ or in $C_{fr}$ (or in both) will move the decision boundary in order to restrict the $R_d$ region (boundary $B'$). Conversely, if $P(\theta_d)$ or $C_{fa}$ (or both) increase then the boundary will move in order to restrict $R_p$. In (b) it is shown that a restriction in $R_d$ (boundary $B'$) corresponds to a lower Bayes threshold $\tau'$ for the LR, and vice-versa. 93

6.6. Probability of false rejection ($P_{fr}$) and false acceptance ($P_{fa}$) and its relationship to $p(E | \theta_p)$ and $p(E | \theta_d)$ respectively. 95

6.7. Elements in the decision process using LR-based forensic automatic speaker recognition. 96

6.8. Logarithmic scoring rule. The x-axis represents the posterior probability of $\theta_p$, which may be viewed as the “forecast” of the fact finder about whether $\theta_p$ is true considering all the available knowledge in the case. 97

6.9. Example of Tippett plots showing the actual LR distributions and the rates of misleading evidence when $\theta_p$ and $\theta_d$ are respectively true. (a) shows the Tippett plots and (b) their corresponding FA/FR rates plot. 100

6.10. Value of $C_M$ (Equation 6.11) for different decision thresholds. (a) GMM-SVM-SuperVector system (calibration is not considered) and (b) Logistic regression fused system (calibration is considered). $C_{fr} = C_{fa} = 1$. Bayes thresholds (Equation 6.15) are shown as vertical lines. 102

6.11. Example of APE plot. 104

6.12. DET curves (a), Tippett plots (b) and $Cl_{lr}$ and APE curves (c) for three simulated systems (System 1, System 2 and System 3). LR values have been randomly generated in order to plot these curves. 106

6.13. Prior entropy with respect to $P(\theta_p)$. 108

6.14. Reduction of uncertainty (information gain) due to evidence analysis, for all possible values of the evidence. The area of the ellipses represent entropy, i.e., uncertainty. 109

6.15. Information change due to evidence analysis and assignment of reference probabilities. The cross-entropy consists of the posterior entropy of the reference (inner ellipse, uncertainty) plus the divergence between the reference and the posterior probability of the system (red darker area in outer ellipse, information loss). 111

6.16. Comparison of the $ECE$ plot (a) to APE plots and $Cl_{lr}$ (b). ATVS SVM-SV system in NIST SRE 2006. A high calibration loss is observed, as no LR computation technique has been performed on the scores. 113

6.17. $ECE$ plots for the individual systems based on SVM-SV (a) and GMM (b), and for the fused system using logistic regression (c). 116

XXII
6.18. Proposed classification of LR computation techniques. Pure LR computation does not change discriminating power, whereas hybrid discrimination-calibration should not degrade discriminating power. ........................................ 117
6.19. (a) FA/FR rate distributions of a real ATVS-UAM speaker recognition system in NIST SRE 2006. (b) Comparison of pure-calibration score to log (LR) transformations. ...................................................... 120
6.20. Comparison of generative hybrid discrimination-calibration LR computation techniques. Left figure shows the pdfs assigned to the data (the scores used are the same as in Figure 6.19). In (a) and (b), Gaussian distributions are assumed, having equal variances in (b). In (c) and (d), Kernel Density Functions with different smoothing parameters are used. ..................................................... 121
6.21. Example of MAP adaptation using synthetic scores. ................................. 123
6.22. $C_{ltv}$ for suspect-adapted LR computation with different amount of suspect scores $M$ and different relevance factors $r$ for the selected subset of the 8conv4w-1conv4w condition in NIST 2005 SRE. ...................................................... 124
6.23. Tippett Plots for suspect-adapted LR computation with different amount of suspect scores $M$ and $r = 1$. .......................................................... 125
6.24. DET plots showing discrimination performance for the ATVS-UAM GMM system for different LR computation methods. .......................... 126
6.25. Tippett plots showing distribution of scores and log_{10} (LR) values for the ATVS-UAM GMM system for different LR computation methods. .................. 127
6.26. APE plots showing discrimination and calibration performance for the ATVS-UAM GMM system with different LR computation methods. .............. 128
6.27. $ECE$ plots showing the discrimination and calibration performance of the ATVS-UAM GMM system using information-theoretical magnitudes, for different LR computation methods. .............................. 129
6.28. The discrimination-presentation level in order to adapt the proposed hierarchical methodology for non-score-based LR computation techniques. .............. 131
7.1. The forensic level in the proposed hierarchical methodology for forensic automatic speaker recognition. .............................................................. 134
7.2. Performance of the LR values computed in the simulated case using the ATVS-UAM GMM system and suspect-adapted evidence evaluation. DET plots (a), Tippett plots (b), APE plots and $C_{ltr}$ (c) and finally $ECE$ plots (d) are shown for the scores of the ATVS-UAM GMM system and after suspect-adapted LR computation. ............................................................ 139
7.3. Performance of the DGHC system in the simulated case using the BDRA database. DET plots (a), Tippett plots (b), APE plots and $C_{ltr}$ (c) and finally $ECE$ plots (d) are shown for the LR values delivered by the forensic automatic speaker recognition system. .......................................................... 142
7.4. A microscopic image of glass fragment.

7.5. A cross-section of car paint observed under an optical microscope (magnification 40x).

7.6. ECE plots for glass evidence. In (a-c) examples of three sets of bivariate LR values are shown, namely \((Na', Si')\), \((K', Na')\) and \((Ca', Na')\). In (d) the ECE plot is shown for the LR for the proposed technique based on graphical models.

7.7. ECE plots for paint evidence. Examples of two sets of univariate and one set of bivariate LR values: TOL (a), MST (b) and (MMA, M2P) (c). In (d), the result for the LR values computed using graphical models is shown.
List of Tables

1.1. Classification of forensic disciplines according to the conclusions of the forensic analysis usually reported (from Champod 2000). ............................................................... 7
1.2. Forensic disciplines apart from DNA where the LR framework has been proposed. ...................................................... 10

2.1. Example of DNA profiles. At the top, a profile using a 10-loci system (SGM Plus 10 loci system used in the UK National DNA Database). Below, a profile using a 13-loci system (core STR markers for the US/FBI Combined DNA Index System - CODIS). 8 loci are shared by both systems. Adapted from Dessimoz and Champod [2007]. .............................................................................................................. 32

3.1. Databases and conditions of NIST SRE corpora. Source: van Leeuwen et al. [2006]. 44
3.2. NIST SRE test conditions since SRE 2004. The core condition for each year appears in boldface. Test conditions used in this Thesis appear in italic (1conv4w-1conv4w and 8conv4w-1conv4w in both SRE 2005 and 2006). ............................................................... 47
3.3. Sample size of NIST SRE test conditions used in this Thesis. ............................................................... 48

5.1. Total number of models $N$ in each T-Norm cohort ............................................................... 75
5.2. Tnorm and KL-Tnorm in GMM system using $K = 75$ models for both techniques (T-Norm values are averaged from 10 random selection trials) ............................................................... 77
5.3. T-Norm and KL-T-Norm in SVM-GLDS system using $K = 75$ models for both techniques (T-Norm values are averaged from 10 random selection trials) ............................................................... 77

6.1. Values of the costs and the prior probabilities in NIST SRE. ............................................................... 100
Chapter 1

Introduction

This PhD Thesis is focused on the use of automatic speaker recognition systems for forensic evaluation of the speech evidence. In particular, the Thesis explores the characteristics of the forensic environment at different levels; develops a novel hierarchical methodology for forensic automatic speaker recognition considering the requirements of forensic science; and reports experimental results in order to illustrate the validity of the proposed methodology as well as the contributions presented.

Forensic science is defined as the application of the sciences as it pertains to legal matters or problems [Gialamas, 2000]. One of the branches of forensic science, namely criminalistics, is the profession and scientific discipline oriented to the recognition, identification, individualization and evaluation of physical evidence by the application of natural science to law-science problems (from Gialamas [2000]). Here, the aim is obtaining the identity of an unknown source from the scientific analysis of the evidence presented in a trial [Champod, 2000; Gialamas, 2000]. Therefore, criminalistics has been dubbed the science of individualization [Champod, 2000], understood as the process which looks for the certainty of not confusing a person or object from the rest of people or objects considered. We will call forensic identification to the process which aims at achieving individualization, according to Champod [2000]. Here, the role of the forensic scientist is the evaluation of the contribution of the forensic evidence to the relevant question for the fact finder, usually “Does the incriminatory material of unknown origin come from a given known source?” Then, it is the duty of the fact finder to establish a final judgment about the identification of the unknown person [Aitken and Taroni, 2004; Dawid, 2004], considering both the weight of the evidence, evaluated by the forensic scientist, and any other information in the case [Aitken and Taroni, 2004; Champod, 2000; Champod and Meuwly, 2000; Evett, 1998; Lindley, 1977]. Figure 1.1 illustrates the information present in such an inferential process.

Unfortunately, it is not a trivial task to include the weight of the evidence in the logical process carried out by the fact finder. In fact, there have been many reported cases in the history of forensic identification where the information about the evidence has been exaggerated, misleading, erroneous or even fallacious [Aitken and Taroni, 2004; Cole, 2005b; Dawid, 2004; DoJ, 2006; Koehler, 1993; Saks and Koehler, 2005]. Moreover, there are many references in the
1. INTRODUCTION

Figure 1.1: Information present in a case in order to answer the relevant question for the fact finder, which is usually “Do the control and recovered material come from the same source?”

literature, most of them supported by rigorous experimental studies, where the importance of forensic reporting is highlighted as a critical factor in order to avoid errors in the interpretation of the evidence which may lead in many cases to erroneous decisions about the identity of the unknown individual in the case [Aitken and Taroni, 2004; Dawid, 2004; Drod et al., 2006; Evett, 1993; Kaye and Koehler, 1991; Koehler, 1993; Saks and Koehler, 2005; Taroni and Aitken, 1998a,b]. It seems therefore essential for any forensic discipline to establish the logical and scientific protocols in order to give meaningful information to the court, avoiding confusing and obscure statements.

In fact, although forensic identification has been systematically performed during the 20th century, especially in classical areas such as fingerprints, ballistics, paint, glass, tool marks, footwear marks, etc.; the debate about the presentation of forensic evidences in a court of law is currently a hot topic in many scientific and legal fora [Champod, 2006; Kennedy, 2003; National Academy of Sciences, 2005; Saks and Koehler, 2005]. One of the main reasons for this discussion arises from the American Daubert rules for the admissibility of scientific evidence in court [U.S. Supreme Court, 1993]. According to these rules, the U.S. Supreme Court suggests that scientifically sound techniques presenting standard procedures and demonstrating their testability, accuracy and acceptance in the scientific community are likely to be admitted in a U.S. federal court of law. On the other hand, non-scientific statements, such as expert testimony lacking of scientific foundations, are likely to be rejected. The implications of these rules are in
accordance to many opinions of forensic experts worldwide [Aitken and Taroni, 2004; Champod, 2006; Champod and Evett, 2001; Cole, 2005a; Curran et al., 2000; Evett, 1998; Kennedy, 2003; Saks and Koehler, 2005; Stoney, 1991], demanding more transparent procedures and a scientific framework for a logical and testable interpretation of the forensic evidence. The debate also considers that existing techniques which have been assumed by the court as error-free are starting to be questioned (see, for example, Cole [2005b] for a complete study regarding latent fingerprint identification). This has been partly due to some critical errors in positive identification reports, highlighted by the mass media, like the Mayfield case in Madrid terrorist attacks in 11 March 2004 (Figure 1.2 [DoJ, 2006; Heath and Bemton, 2004]). Also, forensic case data is not sufficiently integrated into the police investigative processes as it should be, and the use of standard models for crime analysis making use of evidence interpretation are being more and more demanded [Ribaux and Margot, 2003; Ribaux et al., 2006].

Figure 1.2: An example of error in a positive identification. On the left, latent fingerprint believed to belong to a terrorist involved in the train bombings in 11 March 2004 in Madrid, Spain. On the right, a fingerprint resulting from a database search and belonging to Brandon Mayfield from Portland, Oregon. On the basis of these two prints the FBI fingerprint examiners erroneously identified Mayfield as the source of the latent print. 15 characteristic points (minutiae) are marked in both images. Source: DoJ, 2006.

These requirements should be considered in order to use automatic speaker recognition systems as a discipline for forensic identification. Automatic speaker recognition has been subject of study for three decades, as a branch of speech technologies [Bimbot et al., 2004; Campbell, 1997; Gonzalez-Rodriguez et al., 2007c]. However, it has been in the last decade when the area has experimented a greater growth. It is widely accepted that one of the main initiatives for fostering research and technology development in automatic speaker recognition have been due to the efforts of the American National Institute of Standards and Technology (NIST) [NIST]. Due to a continuous series of Speaker Recognition Evaluations (SRE) since 1996 [Przybocki et al., 2007], research has led to a mature technology with known and acceptable performance in
1. INTRODUCTION

many challenging environments [Bimbot et al., 2004; Reynolds, 2003b; van Leeuwen et al., 2006]. The main consequences of this technological push have been a tremendous effort in database acquisition and collection [Campbell and Higgins, Campbell et al., 2004a; ELRA; Garofolo et al., LDC; Ortega-Garcia et al., 2000] and the use of common protocols for automatic speaker recognition systems assessment and presentation of results [Brümmer and du Preez, 2006; Przybocki et al., 2007; van Leeuwen et al., 2006]. Moreover, automatic speaker recognition is now also integrated into biometrics [Jain et al., 2007; Wayman et al., 2005b], a wider group of authentication technologies where the identification of the unknown individual is performed using other traits apart from speech. Thus, automatic speaker recognition has also been fostered by the increasing growth in biometric technologies, evidenced by the publication of reference books [Bolle et al., 2004; Jain et al., 1999, 2007; Müller, 2007; Ratha and Bolle, 2004; Ratha and Govindaraju, 2007; Ross et al., 2006; Zhang, 2002], conferences [Jain and Ratha, 2004; Kittler and Nixon, 2003; Maltoni and Jain, 2004; Odyssee, 1996-2008; Voicebiocon, 2007; Zhang and Jain, 2004], common benchmark tools and evaluation [FVC, 2006; Grother et al., 2003; Maio et al., 2004; Phillips et al., 2000; Przybocki et al., 2007; Sansegundo, 2006; van Leeuwen et al., 2006; Wilson et al., NISTIR 7123; Yeung et al., 2004], cooperative international projects [BioSec, 2004; Biosecure, 2004; COST-275, 2005; Reynolds, 2003c], international consortia [BC, 2003; EBF, 2003; standardization efforts BioAPI, 2002; SC37, 2005], and increasing attention both from government [DoD, 2005] and industry [International Biometric Group, 2006].

This PhD Thesis addresses the use of automatic speaker recognition systems for forensic purposes, which is known as forensic automatic speaker recognition [Alexander, 2005; Dessimoz and Champod, 2007; Drygajlo, 2007; Gonzalez-Rodriguez et al., 2007a; Meuwly, 2001; Ramos-Castro et al., 2006b]. In this introductory chapter we present the basic knowledge about scientific methods in forensic identification. The requirements arisen from the so-called coming paradigm shift in forensic identification are also sketched, from which the motivation of this Thesis is also derived. We finish the chapter by stating the Thesis, giving an outline of the Dissertation, and summarizing the research contributions originated from this work.

Although no special background is assumed in this chapter, the reader will benefit from introductory readings in forensic evidence evaluation such as Lucy [2005] and in automatic speaker recognition such as Bimbot et al. [2004]; Campbell [1997].

1.1. Forensic identification science in the 21st century

There is a well-known principle in forensic science known as Locard’s principle [Locard, 1926], which states that every contact leaves a trace. This can be illustrated as follows: suppose an unknown individual phones a victim in order to blackmail her. A trace may be left by the blackmailer if the police wire-taps the phone call. Forensic transfer evidence, or simply forensic evidence, is defined as the relationship between such trace, whose source is unknown, and some other material which was originated by a known source, both of them related to a given crime or offense [Aitken and Taroni, 2004]. In every case, and according to Locard’s principle, we may
1.1 Forensic identification science in the 21st century

have two types of material: on the one hand, we will have some material of unknown origin, which we will call recovered material, sample or mark [Aitken and Tarori, 2004]. The recovered material may be transferred from the offender to the scene of the crime (e.g., recorded speech, fingermarks, blood stains, fibers from clothes, footwear marks, etc.) or vice-versa (e.g., glass fragments from a window, paint flakes from a wall, fibers from the victim’s clothes, etc.). On the other hand, we may have some material whose origin is known, referred to as the control material [Aitken and Tarori, 2004]. This control material may be obtained from a given suspect (e.g., a blood sample, a fingerprint, a footwear print, a controlled recording, etc.) or from the scene of the crime (a sample of glass from a window, some fibres from the victim’s clothes, etc.). Figure 1.3 illustrates the traces involved in a case where transfer evidence is present.

![Figure 1.3: Transfer evidence.](image)

In forensic identification, the role of the forensic scientist is to examine the material available (recovered and control material) and to evaluate the contribution of these findings with regards to competing propositions arising from the circumstances [Aitken and Tarori, 2004; Dawid, 2004; Evett, 1998]. When the propositions are stated in terms of the source of the presented materials, we will talk about a source attribution problem [Cook et al., 1998], such as the ones considered in this Thesis. Thus, in such case the prosecutor view will suggest that the questioned material and the control material come from the same source (namely \( \theta_p \), the prosecution hypothesis) whereas the defense will support that the source of the questioned material and the control material is not the same (\( \theta_d \), alternate or defense hypothesis) [Champod and Meuwly, 2000; Evett, 1998].

For example, consider the blackmail transfer evidence case described before, with a control recording form a suspect and a recovered questioned utterance from a wire-tap. The aim of a forensic scientist in such a case is reporting meaningful information in order for the court to assess the weight of the forensic evidence in this context of identification of sources [Champod, 2000]. In most cases, the relevant question for the fact finder is:
Do the control and recovered material come from the same source?

In the following sections we will describe classical ways of addressing this question, and we will review the way of giving a proper answer as suggested by the literature.

1.1.1. Classical forensic identification

Classically, there have been two different approaches to forensic reporting regarding source attribution, addressed by areas such as fingerprint, voice, face, signature, DNA, tool marks, paint, glass, fibers or firearms. The first approach has been to provide just “identification” or “exclusion” decisions, which results in a very high percentage of non-reporting cases. This approach has two main drawbacks:

1. **The use of subjective thresholds.** Any system or technique comparing two samples of material is always subjected to uncertainty, especially in forensic conditions. This uncertainty is conditioned by many other factors apart from the forensic evidence being analyzed. Then, if the forensic scientist takes the subjective decision of identification or exclusion, she will be ignoring all other knowledge related to the case, usurping the role of the court in taking this decision. Moreover, “... the use of thresholds is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert” [Champod and Meuwly, 2000], even if these thresholds are adopted from objective measurements. Therefore, an identification/exclusion conclusion constitutes a leap of faith [Stoney, 1991], where the expert jumps from reasoning under uncertainty to absolute conclusions. It is the role of the fact finder, and not the role of the expert, to take such decision.

2. **Non-efficient use of the available knowledge.** The large amount of non-reporting cases that this identification/exclusion process induces may lead to ignoring a significant amount of knowledge about the case. This is a drawback, since “... there is no logical reason to suppress probability statements ... because ... any piece of evidence is relevant if it tends to make the matter which requires proof more or less probable than otherwise” [Champod and Meuwly, 2000].

The second classical approach to forensic reporting in this area consists in the use of a verbal scale of identification probabilities (typically “identification” / “very probable” / “probable” / “not conclusive” / “elimination”), as possible outputs of the forensic analysis. However, this approach falls in the same errors as has just been noted, as it makes use of several subjective thresholds, and again ignores the information relative to every case but not related to the forensic evidence.

Table 1.1 shows several identification-of-the-source forensic disciplines, where a distinction is made between those frequently leading to identification/exclusion decisions and those usually yielding evidence corroborating any of the hypotheses.
Table 1.1: Classification of forensic disciplines according to the conclusions of the forensic analysis usually reported (from Champod 2000).

<table>
<thead>
<tr>
<th>Individualization</th>
<th>Corroborative Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fingerprints</td>
<td>Microtraces (glass, paint, hairs, fibers)</td>
</tr>
<tr>
<td>Footwear marks</td>
<td>Biological fluids (now mostly DNA evidence)</td>
</tr>
<tr>
<td>Earmarks</td>
<td>Drugs and toxicology</td>
</tr>
<tr>
<td>Tool marks and firearms</td>
<td>Explosives and fire residues analysis</td>
</tr>
<tr>
<td>Questioned documents</td>
<td>Soils</td>
</tr>
</tbody>
</table>

1.1.2. The need of a paradigm shift

According to Saks and Koehler [2005], a coming paradigm shift in forensic identification is upcoming, mainly impelled by several scientific and legal forces, identified as:

- Recent changes in the legal admissibility standards for expert testimony.
- The discovery of erroneous convictions and identification decisions.
- Studies about error rates in several forensic identification disciplines.
- DNA as a model for a scientifically sound identification science [Saks and Koehler, 2005].

One of these driving forces, namely the Daubert rules [U.S. Supreme Court, 1993], define a set of requirements for scientific evidence to be accepted in a U.S. federal court of law, and can be summarized as follows:

1. Has the technique been tested in actual field conditions? This empirical validation should be extended to real casework conditions, not just to a synthetic laboratory test. Disciplines not tested in realistic conditions (such as polygraphy) would not meet Daubert criteria.

2. Has the technique been subject to peer review and publication?

3. What is the known or potential accuracy of the technique?

4. Do standards exist for the control of the technique’s operation? Disciplines with different uses in different laboratories (e.g., penile plethysmography) would not meet Daubert criteria.

5. Has the technique been generally accepted within the relevant scientific community? This was the earlier standard for admissibility before Daubert, namely the Frye test.

---

\[1\] The term paradigm shift is understood in Saks and Koehler [2005] and thorough this Dissertation as a shift towards scientific methodology, and not as a Kuhnian paradigm shift [Kuhn, 1962].
Daubert rules, added to the evidence of errors in some well-established forensic areas, have led to reconsider whether the procedures used for forensic interpretation and reporting are indeed scientific or not [Cole, 2005a; Saks and Koehler, 2005]. A need of transparency and testability in the techniques used is demanded in order to report to court results of the assessment of the methodology in use. This is in accordance to the ideas expressed by scientists and experts [Champod and Evett, 2001; Evett, 1998; Gonzalez-Rodriguez et al., 2007b; Meuwly, 2001; Rosd, 2002], lawyers [Cole, 2005a; Kaye and Koehler, 1991; Saks and Koehler, 2005], and statisticians [Aitken and Taroni, 2004; Curran et al., 2000; Dawid, 2004] worldwide. Moreover, it has been demonstrated that no forensic identification discipline is infallible, even considering some well-established disciplines which were considered as error-free in the past (e.g., fingerprints [Cole, 2005b]). These demonstrations have come either from the scientific community [Bonastre et al., 2003; Champod and Evett, 2001; Innocence Project] or from mistakes in real trials [Cole, 2005b; DoJ, 2006; Heath and Bentorn, 2004; Innocence Project]. In this sense, hardly testable statements such as positive identifications justified by the expert’s experience in the field should be avoided [Champod, 2000; Cole, 2005a; Saks and Koehler, 2005].

Unfortunately, forensic evidence has been largely presented in Court in the form of expert opinions based on experience, with conclusions expressed in the form of hard match or non-match statements or making use of verbal scales of probability of identification, given evidence [Champod, 2000; Champod and Meuwly, 2000; Saks and Koehler, 2003; Stoney, 1991]. The process leading from evidence to conclusion is often opaque, either because it lacks scientific rigor and is inherently unfalsifiable, or because the approach is inadequately tested, and thus potential performance or error rates cannot be known. Not surprisingly, this has often resulted in legal discussion about the acceptance of expert testimony.

1.1.2.1. DNA profiling as a scientific model

In order to cope with these emerging requirements, DNA analysis is appointed as the golden standard of scientific forensic evidence [Saks and Koehler, 2005]. DNA profiling was first presented in Court in the 1980s, and in the extremely short period of time since, it has not just become the standard to be emulated, but has also brought into question all other identification-of-the-source forensic disciplines, some of them with a long tradition of expert testimony, such as fingerprints, handwriting or ballistics [Butler, 2005; Saks and Koehler, 2005]. DNA profiling has solid and well-known scientific foundations. Moreover, it is probabilistic: avoiding individualization or exclusion statements for the determination of the source of the evidence, DNA evidence is often presented using frequencies, match probabilities and inclusion or exclusion probabilities [Aitken and Taroni, 2004; Balding, 2005; Butler, 2005; DNA Advisory Board, 2000]. Also, several voices in the forensic science community advocate assessing the weight of the evidence with likelihood ratios (LR) [Aitken and Taroni, 2004; Champod and Meuwly, 2000; Evett, 1991; Evett and Buckleton, 1996], computed using probabilistic and statistic methods found in the literature [Aitken and Taroni, 2004; Balding, 2005]. This LR numerically expresses support for any of the hypotheses involved in a case, and its value can be logically combined with other
judgments related to the rest of background information. In DNA typing, the $LR$ approach allows scrutinizing and inspection by fact finders and forensic scientists [Saks and Koehler, 2005], thus converting DNA typing into a transparent and understandable discipline. Moreover, testability is achieved by reference population data to assess reported performance levels, resulting in known potential and sufficiently low error rates, and widely-accepted methods.

1.1.3. Emulating DNA in forensic automatic speaker recognition

One of the main advantages of the $LR$ approach relies in its universality. It is possible for any forensic discipline to compute $LR$ values by statistically modelling variation in the forensic samples under analysis. Thus, the $LR$ framework has been proposed as an unifying methodology for evidence evaluation [Aitken and Taroni, 2004; Dawid, 2004; Evett and Buckleton, 1996; Lindley, 1977]. In fact, the debate about the coming paradigm shift fosters an emerging interest in the scientific community on the use of statistical methods for forensic identification, evidenced by conferences and seminars [ICFIS, 2008; Institute of Forensic Research, 2007; Joseph Bell Centre, 2005; National Academy of Sciences, 2005], international research working groups [ENFSI] and recently published scientific books [Aitken and Taroni, 2004; Balding, 2005; Curran et al., 2000; Lucy, 2005; Rose, 2002; Taroni et al., 2006]. As a consequence, the $LR$ framework has been proposed for many disciplines in recent years as illustrated in Table 1.2.

This Thesis addresses the use of automatic speaker recognition systems for forensic identification following the procedures derived from the DNA golden standard. Automatic speaker recognition, also known as voice biometrics, is defined as the use of a machine to recognize persons from the speaker voice [Bimbot et al., 2004; Campbell, 1997; Gonzalez-Rodriguez et al., 2007c]. An automatic speaker recognition system is able to compare speech samples, yielding a similarity measure between them. If used in a forensic case involving a control and recovered speech material, this similarity can be viewed as forensic evidence. Therefore, such evidence can be evaluated and assessed following $LR$ methodology derived from DNA.

1.2. Motivation of this PhD Thesis

The motivation of this Thesis is justified by the following observations from the state-of-the-art:

1. In forensic identification there exist different approaches for forensic interpretation of the evidence using biometrics in general and speaker recognition in particular [Champod and Meuwly, 2001; Dessimoz and Champod, 2007; Kinzell, 1994; Nakasone and Beck, 2001; Rose, 2002]. Therefore, it is still needed to stimulate convergence regarding evidence evaluation using automatic speaker recognition systems. According to the coming paradigm shift in forensic identification [Saks and Koehler, 2005], a new scenario should be proposed following the DNA procedures as the golden standard of scientifically sound methods in forensic science. On the one hand, transparency is needed in order to allow both fact
<table>
<thead>
<tr>
<th>Forensic discipline other than DNA</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fingerprints</td>
<td>Champod and Evett [2001]</td>
</tr>
<tr>
<td></td>
<td>Champod [2006]; Neumann et al. [2007]</td>
</tr>
<tr>
<td></td>
<td>Egli et al. [2007]; Neumann et al. [2006]</td>
</tr>
<tr>
<td>Microtraces (glass)</td>
<td>Evett and Buckleton [1990]; Lindley [1977]</td>
</tr>
<tr>
<td></td>
<td>Aitken and Lucy [2004]; Curran [1997]; Curran et al. [2000]</td>
</tr>
<tr>
<td></td>
<td>Aitken et al. [2007]; Ramos et al. [2007]</td>
</tr>
<tr>
<td>Microtraces (paint)</td>
<td>Lindley and Eggleston [1983]; McDermott et al. [1999]</td>
</tr>
<tr>
<td></td>
<td>Ramos et al. [2007]</td>
</tr>
<tr>
<td>Microtraces (fibers)</td>
<td>Aitken and Taroni [2004]; Champod and Taroni [1999]</td>
</tr>
<tr>
<td>Microtraces (hair)</td>
<td>Hoffman [1991]</td>
</tr>
<tr>
<td>Earmarks</td>
<td>Champod et al. [2001]</td>
</tr>
<tr>
<td>Firearms</td>
<td>Bunch [2000]; Champod et al. [2003]</td>
</tr>
<tr>
<td>Explosives</td>
<td>Pierrini et al. [2007]</td>
</tr>
<tr>
<td>Tool marks</td>
<td>Champod et al. [2003]</td>
</tr>
<tr>
<td>Documents</td>
<td>Champod et al. [1999] (Dreyfus case, Darbnx et al., 1908)</td>
</tr>
<tr>
<td>Envelopes</td>
<td>Aitken and Taroni [2004]</td>
</tr>
<tr>
<td>Handwriting</td>
<td>Aitken and Taroni [2004]</td>
</tr>
<tr>
<td>Phonetic-acoustic speaker recognition</td>
<td>Gonzalez-Rodriguez et al. [2007b]; Rose [2002, 2006b]</td>
</tr>
<tr>
<td></td>
<td>Champod and Meuwly [2000]; Khodai-Joopari [2006]</td>
</tr>
<tr>
<td>Biometrics (including automatic speaker recognition)</td>
<td>Champod and Meuwly [2000]; Meuwly [2001]</td>
</tr>
<tr>
<td></td>
<td>Drygailo [2007]; Gonzalez-Rodriguez et al. [2007b]</td>
</tr>
<tr>
<td></td>
<td>Campbell et al. [2005]; Ramos-Castro et al. [2006b]</td>
</tr>
</tbody>
</table>

Table 1.2: Forensic disciplines apart from DNA where the LR framework has been proposed.
finders and forensic scientists to scrutinize each step of the evidence evaluation processes in a scientific way. On the other hand, testability is needed in order to scientifically assess the performance of forensic disciplines, which is demanded by the Daubert rules as well as many experts from the forensic science community.

2. According to Daubert rules, testability in forensic automatic speaker recognition systems should be conditioned to the use of common methodologies for the presentation of testing results. Although successful methods for measuring discrimination [Martin et al., 1997] and rates of misleading evidences (Tippett plots, see Gonzalez-Rodriguez et al. [2006]) have been proposed for forensic automatic speaker recognition, they lack a clear definition of the accuracy of the LR values obtained in a test. Recently, several measures of accuracy have been proposed for LR-based automatic speaker recognition [Brümer and du Preez, 2006], and research is needed in order to adapt them to the requirements of the forensic field. Finally, in a forensic context, intuitiveness and clarity are critical requirements of such evaluation methodology.

3. Although the LR approach has been already proposed for automatic speaker recognition systems [Alexander, 2005; Drygajlo, 2007; Gonzalez-Rodriguez et al., 2006; Meuwly, 2001], variability in the speech signal and data scarcity seriously affect the accuracy of LR computation techniques, as it is shown in Alexander [2005]; Gonzalez-Rodriguez et al. [2006]. Thus, research is needed in order to increase robustness of LR computation for automatic speaker recognition.

4. The state of the art of automatic speaker verification technology presents a limited performance in some scenarios [Bonastre et al., 2003; Reynolds, 2003b], far from the accuracy of forensic techniques such as DNA profiling [Butler, 2005; Evett and Buckleton, 1996]. Therefore, constant research is needed in order to improve speaker recognition systems accuracy if they are going to be used in forensic casework, where the conditions of the speech signal are far from ideal.

The observations described above will be discussed in Chapter 2 in relation to the state of the art related to this Thesis.

1.3. The Thesis

The Thesis developed in this Dissertation can be stated as follows:

The emerging requirements for evidence evaluation and reporting in forensic science can be satisfied for forensic automatic speaker recognition by the use of accurate Likelihood Ratios (LR) within a hierarchical methodology consisting of 3 levels: discrimination, presentation, and forensic.
1. INTRODUCTION

The proposed hierarchical methodology defines three abstraction levels, which will be detailed in Chapter 4. At each one of them, several issues are addressed regarding the evidence evaluation and reporting process, from the evidence evaluation methods, to the assessment framework, the interpretation of results and its presentation in court.

1.4. Outline of the Dissertation

The main objectives of this PhD Thesis are:

1. Reviewing and studying the problem of automatic speaker recognition for forensic evidence evaluation.
2. Identifying all the steps which are needed for the use of an automatic speaker recognition system for forensic identification.
3. Analyzing the requirements of each of the steps and their relationship in order to give a coherent methodology.
4. Defining a methodology for the \( LR \)-based evaluation of the evidence using automatic speaker recognition systems based on the DNA paradigm.
5. Establishing a definition and assessment framework of the \( LR \) accuracy, aiming at clear interpretation of results.
6. Improving the discrimination of automatic speaker recognition technology.
7. Improving the accuracy and robustness of the \( LR \) computation process.
8. Applying the proposed evaluation, interpretation and assessment methodology to forensic speaker recognition problems, either simulating real cases or using databases coming from real police investigations.

The Dissertation is structured according to a traditional complex type [Paltridge, 2002] with background theory, literature review, theoretical and practical methods and three experimental studies in which the methods are applied. The chapter structure is as follows:

- Chapter 1 introduces the topic of forensic identification, the proposed framework for the use of automatic speaker recognition for evidence evaluation, and gives the motivation, outline and contributions of this PhD Thesis.
- Chapter 2 summarizes related works and details the motivations for this Thesis based on these previous works.
- Chapter 3 describes the speech databases, protocols and baseline speaker recognition systems used for the experimental results presented in this Dissertation.
Chapter 4 introduces the methodology for forensic automatic speaker recognition proposed in this Thesis. The whole problem is divided into three abstraction levels, representing the steps needed to achieve transparent and testable forensic automatic speaker recognition. These levels are defined as: 1) the *discrimination* level, where the objective is discriminating between same-source and different-source pairs of speech material; 2) the *presentation* level, where the objective is presenting information in a useful and meaningful way; and 3) the *forensic* level, where issues about reporting and testing protocols are stated according to the upcoming needs in forensic science. The rest of the Dissertation focuses on the description of the proposed levels.

Chapter 5 presents the discrimination level in forensic automatic speaker recognition, where the aim is discriminating between same-source and different-sources pairs of control-recovered speech. This level is sufficient in many applications of speaker recognition technology, such as access control or speaker spotting. Several widely used methods for improving the discrimination in automatic speaker recognition are experimentally studied, namely fusion of systems, session variability compensation and score normalization. A contribution at this level is presented, namely KL-T-Norm, a novel speaker- and test-dependent score normalization technique.

Chapter 6 introduces the presentation level for forensic automatic speaker recognition. The need of adding a meaning to the discriminating score leads to the use of LR-based automatic speaker recognition systems. At this level, a LR value which measures the weight of the evidence is obtained from the score yielded by an automatic speaker recognition system. After a theoretical introduction, the need of measuring the accuracy of the computed LR values is also introduced in this Chapter. A novel assessment technique is presented in order to measure the accuracy of LR values with information-theoretical magnitudes. Then, an experimental comparison of LR computation methods is presented, with a novel contribution, namely suspect-adapted LR computation.

Chapter 7 presents the forensic level of the proposed methodology. At this level, the LR values obtained with the automatic speaker recognition system at the presentation level should be presented in a court according to the requirements derived from the emerging demands of forensic identification science. Thus, this level completes the proposed methodology considering LR assessment and reporting issues, as well as the adequacy to standard protocols in forensic science mainly derived from DNA analysis. Moreover, several experimental simulation of forensic cases are described, both using databases simulating forensic cases and real forensic field data. Issues about the generalization of the proposed methodology for other forensic disciplines are also considered here.

Chapter 8 concludes the Dissertation summarizing the main results obtained and outlining future research lines.
Figure 1.4: Dependence among Dissertation chapters.
1.5 Research contributions

The dependence among the chapters is illustrated in Fig. 1.4. In order to properly follow the experimental chapters of the Dissertation, an elemental background on automatic speaker recognition (Bimbot et al., 2004; Campbell, 1997; Gonzalez-Rodriguez et al., 2007c), and Bayesian statistics (Bernardo and Smith, 1994; Jaynes, 2003) is desired.

Some methods developed in this PhD Thesis are strongly based on popular approaches from the pattern recognition literature. The reader is referred to standard texts for a background on the topic (Duda et al., 2001; Theodoridis and Koutroumbas, 2003). This is specially useful in order to deal with Chapters 5 and 6. It will be useful, especially for Chapter 5, to review basic knowledge about methods for speech processing (Deller et al., 1999; Huang et al., 2001; Quatieri, 2001). Chapters 6 and 7 assume a knowledge of the fundamentals of Bayesian statistics (Bernardo and Smith, 1994; Jaynes, 2003) and Bayesian decision theory (Duda et al., 2001; Theodoridis and Koutroumbas, 2003). Finally, for Chapters 6 and 7 it would be desirable to have an elemental background on statistical analysis of the forensic evidence (Aitken and Tarori, 2004; Balding, 2005; Lucy, 2005) and information theory (Cover and Thomas, 2006; Shannon, 1948).

1.5. Research contributions

The research contributions of this PhD Thesis are the following (some publications are repeated in different items of the list):

- **Literature reviews.**

  1. Forensic evidence evaluation techniques in automatic speaker recognition (Gonzalez-Rodriguez and Ramos, 2007; Gonzalez-Rodriguez et al., 2007b; Ramos-Castro et al., 2006b).
  2. Assessment of forensic speaker recognition systems (Gonzalez-Rodriguez et al., 2006; Gonzalez-Rodriguez and Ramos, 2007; Ramos-Castro et al., 2006b) (IBM Research best student paper award).
  3. New requirements in forensic science (Gonzalez-Rodriguez et al., 2007b; Ramos-Castro et al., 2006b) (IBM Research best student paper award).
  4. Automatic speaker recognition (Gonzalez-Rodriguez et al., 2005b).
  5. Score normalization for robust speaker verification (Ramos-Castro et al., 2005a; Ramos-Castro et al., 2007) (RTTH best article finalist).

- **Theoretical frameworks.**

  1. Theoretical framework for the use of speaker recognition for forensic purposes (Gonzalez-Rodriguez et al., 2007b; Ramos-Castro et al., 2006b) (IBM Research best student paper award).
1. INTRODUCTION

2. Theoretical framework for the use of information theory for the assessment of LR values [Ramos, 2007; Ramos and Gonzalez-Rodriguez, 2007; 2008; Ramos et al., 2007].

**Novel methods.**

3. Novel methods in the assessment of LR values [Ramos, 2007; Ramos and Gonzalez-Rodriguez, 2007; Ramos et al., 2007].
4. Novel methods of score normalization in speaker verification [Ramos-Castro et al., 2005a; Ramos-Castro et al., 2007] (RTTH best article finalist).

**Improvements in speaker recognition discrimination.**


**New techniques in speaker verification.**

1. New methods for the improvement of automatic speaker recognition discriminating power [Lopez-Moreno et al., 2007; Ramos-Castro et al., 2005a; Ramos-Castro et al., 2007] (RTTH best article finalist).

**New experimental studies.**

1. Experimental studies of automatic speaker recognition systems in the proposed methodology for forensic automatic speaker recognition [Gonzalez-Rodriguez and Ramos, 2007; Gonzalez-Rodriguez et al., 2007a,b; Ramos and Gonzalez-Rodriguez, 2007; Ramos-Castro et al., 2006b] (IBM Research best student paper award).
2. Robustness in LR-based evaluation of the evidence [Gonzalez-Rodriguez et al., 2006, 2003a,b, 2004; Ramos-Castro et al., 2005a, 2006a].
3. Calibration loss effects in forensic speaker recognition [Gonzalez-Rodriguez and Ramos, 2007; Gonzalez-Rodriguez et al., 2007a,b; Ramos and Gonzalez-Rodriguez, 2007; Ramos-Castro et al., 2006b] (IBM Research best student paper award).
4. Reports on the ATVS-UAM automatic speaker recognition system with forensic applications at several NIST SRE and at the NFI/TNO Forensic SRE [Gonzalez-Rodriguez et al., 2006, 2003a,b; Gonzalez-Rodriguez and Ramos, 2007; Gonzalez-Rodriguez et al., 2007].
1.5 Research contributions

2003b, 2004, 2007a,b; Lopez-Moreno et al. 2007; Ramos-Castro et al. 2005a,b,c, 2006b; Ramos-Castro et al. 2007 (RTTH best article finalist).

5. Robust score normalization in speaker verification (Ramos-Castro et al. 2005a,d; Ramos-Castro et al. 2007) (RTTH best article finalist).

- Application to other forensic disciplines.

1. Robust evidence evaluation methods in forensic biometric recognition (Gonzalez-Rodriguez et al., 2005a; Ramos-Castro et al., 2005b).

2. Information theoretical evaluation of LR values coming from glass and paint evidences (Ramos, 2007; Ramos and Gonzalez-Rodriguez, 2007; Ramos et al. 2007).

Other contributions so far related to the problem developed in this Thesis but not presented in this Dissertation include:

- Literature reviews.


2. Linguistic and phonotactic techniques for speaker and language recognition (Montero-Asenjo et al. 2006).

- Theoretical frameworks.

1. A theoretical framework for the application of biometric evidences in forensic reporting (Gonzalez-Rodriguez et al. 2005a).

2. A theoretical framework for LR computation using graphical models for different forensic disciplines (glass and paint analysis) (Ramos et al. 2007).

3. A theoretical framework for speaker recognition using traditional and semi-automatic methods and features (Gonzalez-Rodriguez et al. 2007b).

- New methods.

1. A function-based HMM approach to signature verification (Fierrez-Aguilar et al. 2007).


- New systems for speaker and biometric recognition.

1. A function-based signature verification system (Fierrez-Aguilar et al. 2007).
2. Support vector machine regression system (SVM-R) for speaker recognition [Lopez-Moreno et al., 2007].

New experimental studies.

1. Comparison of LR methods for evidence evaluation using different biometric modalities [Gonzalez-Rodriguez et al., 2005a; Ramos-Castro et al., 2005b].
2. Robustness of evidence evaluation techniques for different biometric modalities [Gonzalez-Rodriguez et al., 2005a; Ramos-Castro et al., 2005b].
3. Comparison and sensitivity analysis of support vector machine regression with respect to support vector machine classification [Lopez-Moreno et al., 2007].
4. Sensitivity of HMM-based signature verification to different tuning parameters [Fierrez-Aguilar et al., 2007].
5. Information-theoretical assessment of LR computation using factorized graphical models for forensic glass and paint analysis [Ramos et al., 2007].
6. Comparison of different methods for evidence evaluation using traditional features (diphthongal F-patterns in Australian English) [Gonzalez-Rodriguez et al., 2007b].

Applications.

1. Application of the proposed framework for forensic automatic speaker recognition to other approaches in forensic speaker recognition (traditional and semi-automatic speaker recognition based on diphthongal length and formant frequencies) [Gonzalez-Rodriguez et al., 2007b].
Chapter 2

Related works

This chapter reviews the state of the art in forensic automatic speaker recognition and summarizes previous works related to this PhD Thesis.

The chapter is organized as follows. First we cover the topic of automatic speaker recognition. This continuously growing area of research can be broadly divided into text-dependent and text-independent automatic speaker recognition, being the latter a less constrained task. Text-independent automatic speaker recognition, where the linguistic content in each utterance is unknown a priori, is the most typical case in forensic casework. Therefore, it will be the main focus of this Thesis. Second, we introduce the interpretation of the forensic evidence following the likelihood ratio (LR) methodology used in DNA, describing the first attempts to the application of the LR methodology in automatic speaker recognition. Finally, conclusions are drawn.

Some of the works related to this Thesis and found in the literature will be only cited in this chapter, and further extended in subsequent chapters.

2.1. Automatic speaker recognition

In this section we present the related works on text-independent automatic speaker recognition, showing that the current state of the art reveals a mature technology ready to use in many commercial and forensic applications [Reynolds, 2003b]. Nevertheless, human voice is not only related with personal characteristics, but also with many environmental and sociolinguistic variables, as voice generation is the result of a extremely complex process. Thus, any speech segment acquired from any transmission channel will embed a degraded version of speaker specificities and will be influenced by many contextual variables which are difficult to deal with [Bimbot et al., 2004; Reynolds, 2003b].

2.1.1. Identity information in the speech signal

Speech production is a extremely complex process whose result depends on many variables at different levels, including sociolinguistic factors (e.g. level of education, linguistic context
and dialectal differences) and physiological issues (e.g. vocal tract length and shape or the dynamic configuration of the articulatory organs). These multiple influences will be simultaneously present in each speech act, and some or all of them will contain specificities of the speaker. Hence, we need to clarify and clearly distinguish the different levels and sources of speaker information that we should be able to extract in order to model speaker individualities (Figure 2.1).

The identity levels in the speech signal are configured by the speech production process, which is the subject of study of phoneticians and other areas such as engineering, physics or signal processing [Deller et al., 1999; Huang et al., 2001; Rabiner and Schafer, 1978; Stevens, 2000]. There are two main stages in voice production: i) language generation and ii) speech production; and speaker specificities are introduced in both components. In the field of speaker recognition these two components correspond with which is usually known as high-level (linguistic) and low-level (acoustic) characteristics. Automatic speaker recognition systems will intend to take advantage of the different sources of information available in the speech signal, combining them in the best possible way for every speaker [Brümer et al., 2007; Doddington, 2001; Garcia-Romero et al., 2003; Reynolds, 2003c; Reynolds et al., 2003].

In this section, we will describe approaches for automatic speaker recognition at the following levels, from the lower one (spectral) to the highest one (particular use of the language):

**Short-time spectral level**, or simply **spectral level**. Here, the information about the speaker identity is extracted from the spectrum of the speech signal, analyzed in short-time windows. The spectrum of the speech signal is directly related to the dynamic configuration of the vocal tract, which presents speaker-dependent specificities. This level leads the state-of-the-art in text-independent automatic speaker recognition, and has been the main subject of research in the last decade [Reynolds, 2003b].

**Phonotactic level.** At this level the information about the identity of the speaker is embedded
in the particular use of the phones and syllables and their realizations. Phonotactics is fundamental in language characterization, and therefore these features present high language-dependent variability.

**Prosodic level.** Prosody is the combination of instantaneous energy, intonation, speech rate and unit durations that provides speech with naturalness, full sense, and emotional tone. All these characteristics are speaker-dependent, and may be modeled in order to extract information about the speaker identity.

**Idiolectal level.** This level contains the information about speaker identity in the particular use of the words, which not only depends on the speaker, but in many other sociolinguistic conditions.

### 2.1.2. Applications

Due to the acceptability and accessibility of voice signals, the range of possible applications of automatic speaker recognition is wide compared to other biometric traits. According to Bimbot et al. [2004]; Gonzalez-Rodriguez et al. [2007c]; Reynolds [2003b], we can distinguish three major types of applications which take advantage of the biometric information present in the speech signal:

- **Speaker authentication.** Typical applications included in this category are access control (buildings, facilities, etc.), remote authentication (typically by phone in transactions or access personalization) or background recognition (e.g., natural voice checking).

- **Continuous speaker identification and detection,** e.g. watch-list detection in call centers and switchboards, monitoring of telephone lines and wiretapping, surveillance or detection of a speaker in a speech stream (speaker spotting), speaker screening, etc.

- **Forensic speaker recognition,** which comprises two main applications. On the one hand, automatic speaker recognition systems can be used to present speech as evidence in a court of law [Alexander, 2005; Drygaila, 2007; Gonzalez-Rodriguez et al., 2007b; Künzell, 1994; Meuwly, 2001]. On the other hand, automatic speaker recognition can be used in order to gather intelligence in forensic investigations [Ribaux and Margot, 2003; Ribaux et al., 2006].

Forensic evidence evaluation using automatic speaker recognition systems is the main application covered in this Thesis.

### 2.1.3. Technology

The information encoded in the voice signal fundamentally relies on the linguistic content. Hence, it is not surprising that depending on how the linguistic content is used or controlled, we can distinguish two very different types of speaker recognition technologies with different
potential applications. First, text-dependent technologies [Gonzalez-Rodriguez et al., 2007c], where the user is required to utter an specific key-phrase (e.g., Open, Sesame) or sequence (e.g. a four digit PIN like 1 - 3 - 7 - 9), have been a major subject in biometric access control applications through voice authentication [CAVE; Wagner et al., 2004]. The security level of password-based systems can then be enhanced by speaker authentication. In order to avoid possible theft recordings of passwords, text-dependent systems can ask for random prompts, unexpected to the caller, which cannot be easily replied by an impostor. As they are not frequent in forensic automatic speaker recognition, text-dependent systems will not be described in this Thesis. An overview of text-dependent technologies can be found in Gonzalez-Rodriguez et al. [2007c].

The second group of speaker recognition technologies are those known as text-independent [Bimbot et al., 2004]. They are the most frequent case by far in forensic automatic speaker recognition. In this area, research has experienced a significant increase in the last decade. One of the driving factors of such growth has been the NIST SRE series (Speaker Recognition Evaluations) conducted every since 1996 [Przybicki et al., 2007], with extraordinary progress obtained year by year based in blind evaluation using common databases and protocols. NIST SRE databases and evaluation protocols are described in detail in Chapter 3.

2.1.4. Score-based architecture

Automatic speaker recognition is defined as the use of a machine in order to recognize people from their voices [Campbell, 1997]. In most automatic speaker recognition systems an input speech utterance is compared to an enrolled target speaker model, resulting in a similarity measure between them, also called a similarity score [Bimbot et al., 2004; Campbell, 1997; Gonzalez-Rodriguez et al., 2007c]. The target model is obtained from a set of training speech utterances from a known speaker. The process of computing a score from a speaker model and a test speech utterance is usually called a comparison or trial. The trials may be classified as target and non-target trials depending on whether the training and test speech are respectively generated by the same individual or not. Figure 2.2 illustrates the enrollment and score computation stages.

Classically, the way in which the similarity score is processed by the system defines an operation mode [Bimbot et al., 2004]. In the verification or detection mode, the system has to decide whether the identity of the speaker is the same as a claimed one or not. This output decision is generated by comparing the score generated in a trial to a threshold. This verification mode is typical in biometric applications, and sometimes it is referred to as authentication mode. On the other hand, in identification mode the systems decides either the best or \(k\) best identities from a known database associated to the sample of unknown origin, or a decision that the identity claim is not in the database [Singer and Reynolds, 2004]. The identification mode is out of the scope of this Thesis. Figure 2.3 illustrates the general scheme of a speaker verification system.

\[\text{Through this Dissertation, the word trial will refer either to a legal trial or to the trial in order to obtain a score by the speaker recognition system. In case it is not clear by the context, it will be explicitly highlighted.}\]
2.1 Automatic speaker recognition

Feature Extraction
Enrolled Models
Similarity Score
Normalization
Pre-Processing
Feature Extraction
Enrolled S1 Model

(a)

Sample 1 (S1)

Pre-Processing
Feature Extraction
Enrolled S1 Model

Score Normalization

(b)

Sample 2 (S2)

Figure 2.2: Score-based speaker recognition system: (a) enrollment, (b) score computation.

S1: identity claim

Enrolled S1 Model

Score Computation

Similarity Score S1-S2
Decision Threshold
Accepted or Rejected

S2: unknown source

Figure 2.3: General scheme for a speaker verification system.

A trial is referred to as a target or genuine trial when the identity of the source of the questioned speech is the same as the claimed one, otherwise they are called non-target or impostor trials.

2.1.5. Feature extraction and tokenization

The first step in the design of automatic speaker recognition systems is the extraction of features and tokens which contain information about the identity of a speaker. In this Section, we will briefly show different procedures proposed in the literature in order to extract such elements from the speech signal.

2.1.5.1. Short-term spectral feature extraction

The speech signal does not remain stationary, due to the constant changes in the articulatory system within each speech utterance. However, if we restrict our analysis window to short lengths (typically between 10 and 40 milliseconds), our articulatory system is not able to significantly change in such a short-time window (or frame), obtaining what is usually called a pseudo-
stationary signal. The windowing process is depicted in Figure 2.4.

Those windowed signals can be assumed to come from a specific LTI (linear time-invariant) system for that frame, and then we can perform spectral analysis over this short-term window, obtaining spectral envelopes that change frame by frame [Huang et al., 2001; Rabiner and Schafer, 1978].

From the short-time spectrum in each window, a feature vector will be extracted in order to characterize each frame (Figure 2.4). Linear Predictive Coding (LPC) of speech has proved to be a valid way to compress the spectral envelope in an all-pole model with just 10 to 16 coefficients [Deller et al., 1999; Huang et al., 2001]. However, LPC coefficients are strongly correlated among them, which is an undesirable characteristic. Therefore, cepstrum transform [Deller et al., 1999; Furui, 1981] has been proposed in order to obtain pseudo-orthogonal cepstral coefficients. These cepstral features present several attractive properties, and may be directly derived from LPC (LP Cepstral Coefficients, LPCC). They may also be obtained from a perceptually-based mel-filter spectral analysis (Mel-Frequency based Cepstral Coefficients, MFCC) [Deller et al., 1999]. LPCC and MFCC are the most popular feature extraction techniques in speaker recognition, but some other feature extraction techniques are described in the literature, such as PLP (Perceptually based Linear Prediction), RASTA-PLP [Hermansky et al., 1985] or LSF (Line Spectral Frequencies) [Itakura, 1975]. In order to take co-articulation into account, velocity (∆) and acceleration (∆∆) coefficients may be obtained from the static window-based information. This ∆ and ∆∆ coefficients model the speed and acceleration of the variation of frames across adjacent windows.

Figure 2.4: Short-term feature extraction.
2.1.5.2. High-level tokenization

Tokenization is the translation from sampled speech into a time-aligned sequence of linguistic units, or tokens. Hidden Markov Models (HMM) are widely used for phonetic, syllable and word tokenization. HMM as used in speech processing are finite state machines which model the temporal dependency of spectral feature vectors in a probabilistic way [Huang et al., 2001; Rabiner, 1989]. The performance of the tokenization may be improved by the use of language models, which impose some linguistic or grammatical constraints on the high number of combinations of all possible units (phones or words) [Huang et al., 2001].

Basic prosodic features as pitch and energy are also obtained at a frame level. The instantaneous pitch can be determined by, e.g., autocorrelation or cepstral-decomposition based methods, usually smoothed with some time filtering [Rabiner and Schafer, 1978]. Other important prosodic features are those related to linguistic units duration, speech rate, and all those related with accent. In all those cases, precise segmentation is required [Huang et al., 2001; Toledano et al., 2003], i.e., determination of the points in the speech signal where each unit occurs.

2.1.6. Text-independent speaker recognition

Text-independent speaker recognition has been largely dominated by short term spectral-based systems. Since 2000, higher level systems started to be developed with promising results in the same highly challenging benchmarks [Doddington, 2001; NIST; Reynolds, 2003c]. However, spectral systems have continued to outperform high-level systems. In this Section, we describe the most popular techniques for text-independent automatic speaker recognition, both at the spectral and higher levels.

2.1.6.1. Short-time spectral systems

Modeling the short-term spectral features is strongly related to characterizing the different “sounds” that a person is able to produce, due to his/her own vocal tract and articulatory organs [Huang et al., 2001; Rabiner and Schafer, 1978]. As humans need multiple sounds (or acoustically different symbols) to speak in any language, we are clearly facing a multi-class space of characteristics.

Vector Quantization (VQ) techniques are efficient in such multi-class problems [Kinnunen et al., 2006]. At the enrollment stage, a specific VQ model per speaker from training data is obtained. This model quantifies the feature space according to a set of codewords or codevectors, which is called a codebook. The score is then computed as the weighted sum of the minimum distances per frame to the closest codeword in the codebook.

The use of boundaries and centroids instead of probability densities yields poorer performance for VQ than for fully-connected continuous density HMM, known as ergodic HMM (E-HMM) [Matsui and Furui, 1992]. The probability distributions in each state of an E-HMM are typically modelled by a mixture (weighted sum) of Gaussian distributions. Thus, the complexity
of an E-HMM is the product of the number of states times the number of Gaussians per state. It was observed that the performance of E-HMM in speaker recognition was comparable if the complexity remained constant. As a result, a 5-state 4-Gaussian per state E-HMM system will perform similarly than a 4-state 5-Gaussian per state, a 2-state 10-Gaussian per state, or even a 1-state 20-Gaussian per state system. These 1-state HMM are known as Gaussian Mixture Models (GMM) [Duda et al., 2001; Reynolds and Rose, 1995], which is a generative technique where a mixture of multidimensional Gaussians model the probability distribution of the speaker training features. A sample pdf of a GMM is shown in Figure 2.5.

The main advantage of GMM with respect to E-HMM relies on its simplicity, which accelerates computation with no performance degradation. GMM became the state-of-the-art technique in the 1990’s. Different training strategies in order to obtain the GMM pdf have been proposed, such as maximum likelihood (ML) through Expectation-Maximization (EM) or Maximum Mutual Information (MMI) for discriminative training [Duda et al., 2001; Theodoridis and Koutroumbas, 2003]. However, it was the use of Maximum A Posteriori (MAP) adaptation of the means from a Universal Background Model (UBM) which gave GMM a major advantage over other techniques [Reynolds et al., 2000].

Discriminative techniques for speaker recognition such as Artificial Neural Networks have been also used for years [Farrell et al., 1994], but their performance never approached that of GMM. However, Support Vector Machines (SVM) classifiers [Vapnik, 1999] has given GMM its major competitor, and SVM-based systems currently obtain equivalent performance to GMM.
2.1 Automatic speaker recognition systems. First approaches \cite{Campbell:2002} performed score computation from expansions of speech utterances to a high dimensional space. This was achieved by using kernels such as Generalized Linear Discriminant Sequence (GLDS) kernel \cite{Campbell:2002}. Recently, a hybrid technique has been proposed \cite{Campbell:2006}. This approach trains a GMM for every speech utterance. Each GMM is then considered a point in a supervector space of high dimension. A SVM is then used for score computation. GMM-SVM-SuperVector systems (GMM-SVM-SV) present a comparable performance to GMM, outperforming SVM-GLDS systems.

In \cite{Stolcke:2005}, it is demonstrated that adaptation parameters derived from Large Vocabulary Continuous Speech Recognition systems (LVCSR) are excellent features for speaker recognition, being easily modeled as supervectors. Maximum Likelihood Linear Regression was used in \cite{Stolcke:2005} in order to generate this so-called MLLR-supervectors, which were classified by SVM with excellent performance (leading to a MLLR-SVM system). Constrained MLLR (CMLLR) was also used in \cite{Brummer:2007} in a CMLLR-SVM configuration.

2.1.6.2. High-level systems

Short-term spectral features dominated text-independent speaker recognition systems until the work of \cite{Doddington:2001} opened new lines of research. Doddington realized and proved that speech from different speakers differ not only on the short-term spectrum, but also on other characteristics like the word usage. In particular, \cite{Doddington:2001} modeled the word usage of each particular speaker using an n-gram \cite{Huang:2001} which characterized word sequences and their probabilities of appearance. He also demonstrated that those models improve the performance of a baseline spectral GMM system. More important than this particular result is the fact that this work boosted research in the use of higher levels of information (idiolectal, phonotactic, prosodic, etc.) for text-independent speaker recognition. After the publication of \cite{Doddington:2001} a number of researchers met at the summer workshop SuperSID \cite{Reynolds:2003}, where these ideas were further developed and tested on a common benchmark. In this Section we describe two of the most successful approaches exploiting higher levels of information: phonotactic systems, which try to model pronunciation idiosyncrasies, and prosodic systems, which model speaker-specific prosodic patterns.

**Phonotactic systems.** A typical phonotactic speaker recognition system consists of two main building blocks: the phonetic decoders (tokenizers), which transform speech into a sequence of phonetic labels and the n-gram statistical language modeling stage, which models the frequencies of phones and phone sequences for each particular speaker. The phonetic decoders (typically HMM-based) can either be taken from a preexisting speech recognizer or trained ad hoc. For the purpose of speaker recognition, it is not very important to have very accurate phonetic decoders and it is not even important to have a phonetic decoder in the language of the speakers to be recognized. This somewhat surprising fact has been analyzed by \cite{Toledano:2005} showing that speaker-dependent phonetic errors made by the decoder seem to be speaker specific, and therefore they convey useful information
for speaker recognition as long as they are consistent for each particular speaker.

Once a phonetic decoder is available, the tokens (phones) of many sentences from many different speakers can be used to train a Universal Background Phone Model (UBPM) representing all the possible speakers. A speaker phone model \( SPM_i \) is trained using several phonetic decoders of each particular speaker. Since the speech available to train a speaker model is often limited, speaker models are interpolated with the UBPM in order to increase robustness against token scarcity \([\text{Baker et al.}, 2004]\). Once the statistical language models are trained, the procedure to compute a score for a test utterance against a speaker model \( SPM_i \) is represented in Figure 2.6.

The first step is producing the phonetic decoding \( X \) of the test utterance, in the same way as the decodings used to train \( SPM_i \) and UBPM. Then, the phonetic decoding of the test utterance, \( X \), and the statistical models \( (SPM_i, UBPM) \) are used to compute the likelihoods of the phonetic decoding, \( X \), given the speaker model \( SPM_i \) and the background model UBPM. The recognition score is the logarithm of the ratio of both likelihoods. This process, which is usually described as Phone Recognition followed by Language Modeling (PRLM) may be repeated for different phonetic decoders (e.g., different languages or complexities) and the different recognition scores fused for better performance, yielding a method known as Parallel PRLM or PPRLM \([\text{Andrews et al.}, 2002]\).

Recently, several improvements have been proposed on the baseline PPRLM systems. One of the most important in terms of performance improvement is the use of the whole phone recognition lattice \([\text{Hatch et al.}, 2005]\) instead of the one-best decoding hypothesis. The recognition lattice is a directed acyclic graph containing the most likely hypotheses along

---

**Figure 2.6:** Score computation in phonotactic speaker recognition.
2.1 Automatic speaker recognition

Prosodic systems. One of the pioneering and most successful prosodic systems in speaker recognition is the work of Adami et al. [2003]. The system consists of two main building blocks: i) the prosodic tokenizer, which analyzes the prosody, and represents it as a sequence of prosodic labels or tokens; and ii) the n-gram statistical language modeling stage, which models the frequencies of prosodic tokens and their sequences for each particular speaker.

ATVS-UAM has implemented a prosodic system based on Adami et al. [2003] in which the second block is exactly the same for phonotactic and prosodic speaker recognition with only minor adjustments to improve performance (Figure 2.6). The tokenization process consists of two stages. First, for each speech utterance, temporal trajectories of the prosodic features, (fundamental frequency -or pitch- and energy) are extracted. Second, both contours are segmented and labeled by means of a slope quantization process. Figure 2.7 shows a table containing 17 prosodic tokens. One token represents unvoiced segments, while 16 are used for representing voiced segments depending on the slope (fast-rising, slow-rising, fast-falling, slow-falling) of the energy and pitch. Figure 2.7 also shows an example utterance segmented and labeled using these prosodic tokens. Details about the ATVS-UAM prosodic system can be found in Gonzalez-Rodriguez et al. [2007a].

Some other possibilities for modeling the prosodic information that have also proved to be quite successful are the use of Non-uniform Extraction Region Features (NERFs) delimited by long-enough pauses Kajarekar et al. [2004] or NERFs defined by the syllabic structure of the sentence (SNERFs) Shriberg et al. [2004].
2.2. \textit{LR–based analysis of the evidence}

In this section, we present the basics of the likelihood ratio (\textit{LR}) methodology for evidence evaluation. This description will be significantly extended in Chapter 6.

2.2.1. Historical review

The \textit{LR} approach for evidence evaluation was early proposed by Darboux et al. [1908] in the context of the Dreyfus case in 1894 [Aitken and Taroni, 2004; Champod et al., 1999]. At that time, an officer of the French army (Dreyfus) was accused of selling military secrets to the German. The evidence presented was related to document analysis, and probabilistic reasoning was used in a first trial by forensic scientists in order to conclude that the probability of guilt for Dreyfus was 0.9984. This conclusion was wrong because of an error on probability inference.

A retrial occurred in order to discuss the correctness of the previous approach. Outstandingly, Darboux, Appell, and Poincaré, mathematicians members of the French Academy of Sciences, offered their opinions. Among other conclusions, a fundamental fact was pointed out in their reports: a probability of an identification cannot be stated only by the analysis of the evidence, since there is other prior information which must be analyzed and which is not always available for the forensic scientist. Such idea is discussed in the following paragraph extracted from Poincaré [1992]:

\begin{quote}
[...\textit{an effect may be the product of either a cause A or a cause B. The effect has already been observed. One wants to know that probability that it is the result of cause A; this is the a posteriori probability for the cause producing the effect. I want to speak of the probability of this eventually for one who has never observed the result.}]
\end{quote}

However, Darboux et al. [1908] point out that

\begin{quote}
[...\textit{since it is absolutely impossible for us [the forensic experts] to know the a priori probability, we cannot say: this coincidence proves that the ratio of the forgery’s probability to the inverse probability is a real value. We can only say: following the observation of this coincidence, this ratio becomes X times greater than before the observation.}]
\end{quote}

These ideas have been further highlighted and extended during history in relation to disciplines like mathematics [de Finetti, 1930; Olkin, 1958], philosophy [Jeffrey, 1975], forensic science [Kingston and Kirk, 1964; Parker and Holford, 1968] and law [Robertson and Vignaux, 1992]. All this work remarks the importance of the \textit{LR} approach for evidence evaluation. For a detailed historical review see Aitken and Taroni [2004]; Taroni et al. [1998].
2.2.2. The LR approach for evidence evaluation

In a forensic case where some evidence should be evaluated, the hypotheses, defined at the source level [Cook et al., 1998], are usually the following:

- \( \theta_p \) (the recovered and control samples have the same source)
- \( \theta_d \) (the source of the recovered sample is not the same as the source of the control sample, but another one from a relevant population of sources)

According to Darboux et al. [1908], the \textit{a posteriori} or \textit{posterior} probability in favor of any of the hypothesis considers both the prior judgement based on the background information about the case, and the analysis of the forensic evidence. The relationship of all these elements is given by the Bayes’ theorem:

\[
P(\theta_p | E, I) = LR \cdot \frac{P(\theta_p | I)}{P(\theta_d | I)}
\]

where:

\[
LR = \frac{p(e | \theta_p, I)}{p(e | \theta_d, I)}_{e=E}
\]

Here, \( E \) represents the evidence and \( I \) is the background information about the case. Hence, it is shown that the prior probabilities about the hypothesis before evidence analysis are related to the posterior probabilities after evidence analysis by a likelihood ratio (LR), defined in Equation 2.2. Figure 2.8 shows an example which illustrates this Bayesian inference process. The quotient of complementary probabilities in Equation 2.1 are defined as the \textit{odds}, as a way of stating opinions in a binary problem (odds 1 over 4 in favor of the prosecutor hypothesis \( \theta_p \) means also that the odds are 4 versus 1 for the defense hypothesis \( \theta_d \)). Thus, odds are represented as the relationship of the weight in a balance. The prior odds are stated by fact finders with knowledge of the background information in the case, previously to evidence evaluation. With the analysis of the weight of the evidence, expressed in the form of a likelihood ratio (LR), the fact finder may infer the posterior odds in favor of the hypotheses. Thus, the opinion of the fact finder changes according to the weight of the evidence, expressed in the form of a LR.

2.2.3. LR computation in forensic science

The computation of LR values in forensic sciences has been largely influenced by DNA profiling [Aitken and Taroni, 2004; Balding, 2005; Butler, 2005]. DNA profiles are records of the discrete values of short-tandem repetitions (alleles) at given places (loci) in the DNA chain. Table 2.1 shows examples of DNA profiles.

Forensic LR values are obtained relating the similarity between the test and control samples to the typicality in the population of interest of the test and control samples under evaluation [Aitken and Taroni, 2004; Balding, 2005; Curran et al., 2000]. DNA deals with discrete probabilities, and when single stain procedures are considered, if a match is found, the numerator of
Figure 2.8: Bayesian inferential framework in LR-based evidence analysis.

The LR is typically assumed to be one. The denominator is obtained from the random match probability in the population of interest, obtained from a database of DNA profiles. There are also many different considerations regarding assumptions of relevance, corruption of samples, more than one traces and mixtures of genetic profiles, which can help on accurately obtaining the LR values in more complex problems. As this Thesis is focused on continuous LR computation, we will not give an extensive description of the techniques used in discrete LR computation, such as in DNA. The reader can refer to the abundant literature covering the topic, including [Aitken and Tarori, 2004; Balding, 2005; Butler, 2005; Lucy, 2005; Robertson and Vignaux, 1995].

However, most identification-of-the-source disciplines deal with continuously valued measurements or features obtained from control and recovered samples. In order to handle continuous LR computation, several models have been proposed in the literature. [Lindley, 1977] proposed...
a method in order to assign probability distributions to measurements obtained from glass fragments. The proposed model considered two levels of variation, namely within-source level and between-source level, and was later generalized to multiple variables by Aitken and Lucy [2004]. From the LR formula, the following model can be derived [Aitken and Taroni, 2004; Lindley, 1977]:

$$LR = \frac{p(x, y | \theta_p, I)}{p(x, y | \theta_d, I)} = \frac{\int p(x, y | \mu, \theta_p, I) p(\mu | \theta_p, I) d\mu}{\int p(x | \mu, \theta_d, I) p(\mu | \theta_d, I) d\mu \int p(y | \mu, \theta_d, I) p(\mu | \theta_d, I) d\mu}$$

In Equation 2.3, the evidence \(E\) is represented by vectors of features \((x, y)\) where \(x\) are features extracted from the control sample and \(y\) are features extracted from the recovered sample. The parameter \(\mu\) represents the mean of the measurements of \(x\) and \(y\) in each case, and it is integrated out in all cases in Equation 2.3. Thus, the final LR value is independent on the value of the parameter \(\mu\). Lindley [1977] used Gaussian distributions in order to model both between-source and within-sources variations, but different approaches include the use of Kernel Density Functions (KDF) [Aitken and Taroni, 2004], multivariate normal distributions [Aitken and Lucy, 2004; Aitken and Taroni, 2004] and multivariate KDF [Aitken and Lucy, 2004; Aitken and Taroni, 2004; Rose, 2006a].

Recently, the use of graphical models have been proposed in order to handle complex evidence evaluation problems. Undirected graphs have been used for reducing the dimensionality of multivariate LR computation [Aitken et al., 2007; Ramos et al., 2007], leading to low-dimensional problems, much easier to model. On the other hand, complex inference problems have been modeled using Bayesian networks [Taroni et al., 2006], which are graphical models of probabilistic dependence. Bayesian networks allow modeling such problems in a visual way, clarifying the probabilistic inference relationship and giving a deeper insight to the process.

Regarding forensic speaker recognition, in the past it was largely dominated by auditory, phonetic and linguistic analysis singly or in combination [Künzel, 1994; Rose, 2002], although the present consensus among traditional practitioners is that both auditory and acoustic analysis are required [Rose, 2002]. Each of the features used (termed traditional because they relate to features used in traditional phonetic and linguistic analysis [Rose, 2002]) is known to vary across different speakers but also within speakers. Hence, a LR can easily be obtained for each feature if adequate measures of variation among reference populations are available, and measures of variation within the suspect can be extracted from control recordings.

2.3. **LR computation using automatic speaker recognition systems**

In this section, a review of the first attempts for LR computation in automatic speaker recognition is presented. These techniques will be further described and extended in detail in Chapter 6.
2.3.1. First approaches to $LR$–based forensic automatic speaker recognition

The first application of the $LR$ computation methodology to automatic speaker recognition was proposed by Meuwly [2001] for a GMM-based system. There, the scores of an automatic speaker recognition system were used in order to model likelihoods in Equation 2.2. The pdf $p(E|\theta_d)$ in the denominator, namely the between-source pdf, models scores assuming that $\theta_d$ is true. Thus, non-target scores were obtained comparing the questioned speech under analysis with the relevant population of individuals. On the other hand, the within-source distribution, namely $p(E|\theta_p)$ was assigned to scores assuming that $\theta_p$ is true. These within-source target scores were obtained comparing different utterances from the control speech material. This $LR$ computation approach proposed by Meuwly [2001] is summarized in Figures 2.9 and 2.10.

In this approach, the evidence $E$ is the comparison of the recovered speech of questioned source and the control speech from a known suspect. Thus, for automatic speaker recognition systems $E$ will be a similarity score between the recovered and control materials. However, other kind of meta-information (such as signal to noise ratio, transmission channels, subjective quality of the speech signal, etc.) may be also used in order to compute this $LR$ value [Campbell et al., 2005], as we will discuss later.

The information provided by the circumstances of the case regarding the recovered recording leads to a specification of the initial reference population of relevant speakers (potential population). Hence, this approach needs two databases for the calculation and the interpretation of the evidence: the potential population database (P), and the control speech database (C).

The relevant population database (P) is used for modeling the variability of the scores of all the potential relevant sources, using the automatic speaker recognition system. P typically consists of a set of speech utterances from a population of non-suspected individuals. The P database then allows the evaluation of the between-source pdf $p(e|\theta_d)$ given the recovered recording, which will be used in the denominator of the $LR$ (Equation 2.2). Ideally, the conditions of the recordings in the P database should match those of the control speech used in order to model the speaker [Alexander, 2005].

The control speech database (C) is essentially the control speech material, and it is necessary to model the within-source variability. Comparisons performed using different utterances from the control speech database will generate target scores, with which the within-source distribution $p(e|\theta_p)$ will be obtained.

Some authors [Alexander, 2005; Drygajlo, 2007; Meuwly, 2001] identify a third database (the reference database R), which is essentially a subcorpus of the control database C as defined here. While the R database as defined by Meuwly [2001] should contain speech whose conditions are adapted to the P database, the rest of the C database should contain speech whose conditions are adapted to the recovered material. That would make within-source target scores generated with one utterance from the C database and the other one from the R database to be ideally compared in the same conditions as the control-recovered score.
2.3 LR computation using automatic speaker recognition systems

\[ p(e|\theta_p) \]

\[ p(e|\theta_d) \]

\[ e - E \]

\[ e \]

\[ \log (LR) \]

\[ e = E \]

\[ e \]

**Figure 2.9:** LR computation following the methodology proposed by Meuwly [2001] for a GMM-based automatic speaker recognition system. The recovered speech material is referred to as the trace here. Source: Gonzalez-Rodriguez et al. [2006].

**Figure 2.10:** Generative LR computation. In (a), the within- and between-source distributions \( p(e|\theta_p) \) and \( p(e|\theta_d) \) respectively are obtained from the automatic speaker recognition system scores according to Figure 2.9. The LR will be the quotient of such distributions at a given value \( E \) of the evidence \( e \). \( E \) is the value of the similarity score given by the system when comparing the control and recovered materials. In (b), the score to \( \log(LR) \) mapping function is shown. As the variances of the Gaussian distributions in (a) are equal, the mapping in (b) is linear, but it will be quadratic in general for Gaussian pdfs with different variances.
2.3.2. Robustness in LR–based forensic automatic speaker recognition

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 recovered, control and relevant population speech, as well as lack of data for the obtention of accurate distributions [Alexander, 2005; Gonzalez-Rodriguez et al., 2006; Nakason, 2003].

The problem of between-source variability computation is related to the selection and number of available speech utterances from individuals in the relevant population. Given that between-source variability distribution represents the distribution of the evidence score in the relevant population, mismatches in the conditions of population speech and control-recovered speech may lead to inaccuracies in the assignment of the between-source pdf, because inter-session variability seriously affects the behavior and performance of the scores [Kenny et al, 2007]. This effect is illustrated in Figure 2.11 using the ATVS-UAM GMM system in NIST SRE 2006. Hence, there are many problems to be solved when such a matched relevant population database is not available [Alexander, 2005; Gonzalez-Rodriguez et al, 2006].

![Variability of population scores depending on the transmission channel of the speech from the population. Each plot shows the Gaussian fit of the distributions of scores obtained with a recovered speech sample and two populations: a landline telephone population (solid curve) and a cellular telephone population (dashed curve). Scores obtained with the ATVS-UAM GMM system in NIST SRE 2006. Population data extracted from NIST SRE 2004 database.](image-url)
2.3 LR computation using automatic speaker recognition systems

Gonzalez-Rodriguez et al. [2003a] present a study of the effects of session variability between the population and the control-recovered speech, confirming that the mismatches in the population conditions with respect to the control-recovered speech pair lead to a significant degradation in the performance of the LR values. There are also some other works which aim at compensating from session variability at the generative LR computation level. In Alexander [2005]; Alexander et al. [2004], several techniques are presented which model the variability of the scores obtained from a population due to variations in channel conditions. Results show that, if properly modeled, the normalization of the score distribution considering channel variability leads to a significant improvement in the performance of the LR values computed.

Another important source of problems comes from speech data scarcity. It is possible in forensic investigations dealing with voice that the amount of recovered and control speech available to the forensic expert will be highly scarce [Dessimoz and Champod, 2007; Gonzalez-Rodriguez et al., 2006]. In the limit, if just a single control and a single recovered speech utterances are available, it is not always possible to evaluate the within-source variability of the suspect. Therefore, since this is a recurrent problem in forensic automatic speaker recognition, it is necessary to define an interpretation framework for evaluating the evidence even when the control speech material is scarce [Gonzalez-Rodriguez et al., 2006]. Figure 2.12 illustrates how scarcity on target scores obtained with the control speech material may affect the obtention of the within-source pdf.

In Botti et al. [2004], a robust technique is proposed assuming that an accurate model of the within-source distribution for a given suspect can be obtained using target scores from different individuals in the same conditions. However, it has been shown that, even in the same conditions, the target scores coming from different speakers may present very different behavior [Doddington et al., 1998]. Therefore, accuracy in within-source estimation may be improved by exploiting suspect-specific scores.

In Gonzalez-Rodriguez et al. [2006, 2003a] a different approach is proposed, namely Within-source Degradation Prediction (WDP). This technique combines control target scores with between-source distribution knowledge to predict score variability not present in the control data. Experiments presented in Gonzalez-Rodriguez et al. [2006] shows excellent performance when limited control speech material is available. However, WDP aims at fixing the within-source distribution without considering the actual (and unknown) control target scores it claims to represent. Therefore, the predicted within-source pdf will incur in inaccuracies in the within-source distribution obtained.

In this Thesis we propose a generative LR computation technique which is robust to control speech data scarcity. Our motivation comes from the drawbacks of previous works Botti et al. [2004, Gonzalez-Rodriguez et al., 2006]. This novel technique will be presented in Chapter 6.
2.3.3. Discriminative LR computation in forensic automatic speaker recognition

LR computation in automatic speaker recognition was firstly based on the work by Meuwly [2001], which used generative techniques. However, recent techniques have been proposed in order to obtain a LR value according to different discriminative methods [Brümer et al., 2007; Campbell et al., 2005]. Thus, the score from a speaker recognition system can be transformed into a LR value in a discriminative way, as Meuwly [2001] obtained their LR value by generative means. The process of LR computation has essentially the same aim in both cases. Figure 2.13 shows this relationship.

2.4. Chapter summary and conclusions

In this chapter we have summarized several works related to this PhD Thesis. We have started reviewing common techniques for automatic speaker recognition, focusing on text-independent methods. The likelihood ratio (LR) methodology for evidence analysis has been
Figure 2.13: Score-based LR computation. First, a score is computed. Second, a LR is obtained from the score, the control recordings (C database) and the speech from the relevant population (P database). This scheme is common for generative and discriminative LR computation techniques.

then introduced. Finally, LR computation in automatic speaker recognition has been introduced, describing the first approaches and current challenges.

Some of the methods presented in this chapter will be subsequently extended in the rest of the Dissertation.
Chapter 3

Experimental framework

This chapter introduces the experimental framework used in this Thesis, including speech databases, experimental protocols and baseline systems used for presenting results.

The chapter is organized as follows. We first review generalities about evaluation methodologies and statistical significance of experimental results. We then describe the experimental framework followed in this Thesis, mainly based in NIST Speaker Recognition Evaluation (SRE) benchmark tests. A detailed description of the NIST SRE evaluation protocol is provided, as well as an overview of the databases used, focusing in the experimental setup used in this Dissertation. The baseline automatic speaker recognition systems developed at ATVS - Biometric Recognition Group are then detailed. Finally, conclusions are extracted.

3.1. Evaluation methodology

The practice in first research works on automatic speaker recognition as well as other biometrics starting over three decades ago was to report experimental results using biometric data specifically acquired for the experiment at hand [Atal, 1976; Kanade, 1973; Nagel and Rosenfeld, 1977]. This approach made very difficult the fair comparison of different recognition strategies, as the biometric data was not made publicly available.

With the popularity of automatic speaker recognition systems and the creation of new research groups working in the same topics, the need for common performance benchmarks was recognized early in the past decade. This is evidenced by the work on database collection performed by the Linguistic Data Consortium [LDC] and the European Language Resources Agency [ELRA]. Such significant efforts have led to the existence of a wide offer of publicly available speech data common to research groups for comparing results.

One important milestone in speaker recognition technology came in 1996 with the first international Speaker Recognition Evaluation (SRE) conducted by the National Institute of Standards and Technology (NIST), which has been followed by yearly editions until 2006 [Przybocki et al., 2007; van Leeuwen et al., 2006]. In such evaluations, participants gain access to a given
3. EXPERIMENTAL FRAMEWORK

speech database of unknown origin, and a common experimental protocol is established in order to test all systems from participants in a blind and fair way.

Some examples of international evaluations in automatic speaker recognition as well as other biometric traits include the following campaigns: previously cited NIST SRE, NFI-TNO forensic speaker recognition evaluation, organized in 2003 by the Netherlands Forensic Institute and TNO human factors [van Leeuwen and Bouten, 2004; van Leeuwen et al., 2006]; NIST Facial Recognition Technology Evaluations (FERET), starting in 1994 [Phillips et al., 2000b]; NIST Iris Challenge Evaluations (ICE), first organized in 2005 [Phillips, 2006]; Fingerprint Verification Competitions (FVC), held biannually since 2000 [Cappelli et al., 2006]; Signature Verification Competition (SVC), organized in 2004 [Yeung et al., 2004]; and Multimodal Biometrics competition, organized by the Biosecure European Network of Excellence (NoE) during 2007 [BMEC, 2007]. Comparative evaluations of commercial biometric technologies can also be found nowadays by standards institutions like NIST [Grother et al., 2003; Wilson et al., NISTIR 7123] and CESG [Mansfield et al., 2001], or consulting firms like the International Biometric Group [2006].

In order to empirically assess the performance of an automatic speaker recognition or biometric system, a test can be performed from an evaluation database where the identity of each speech utterance is known. Thus, we obtain a set of target scores, where \( \theta_p \) is true, and a set of non-target scores, for which \( \theta_d \) is true.

In order to show the adequacy of the proposed algorithmic contributions, the methodology for evaluation presented in this Thesis will be based on technology evaluation, as it happens in NIST SRE [Phillips et al., 2000a; Przybocki et al., 2007]. The goal of a technology evaluation is to compare competing algorithms thus identifying the most promising recognition approaches and tracking the state-of-the-art. Thus, the novel approaches proposed in this Thesis are always compared to a baseline system. Testing of all algorithms is carried out on a standardized database such as the ones used in NIST SRE. Because the database is fixed, the results of technology tests are repeatable. Some important aspects of a given database are: 1) Number of individuals in the database, 2) number of recording sessions, and 3) number of different samples per session. Most standardized benchmarks in biometrics are technology evaluations conducted by independent groups or standards institutions [Maio et al., 2004; Phillips et al., 2000a; Przybocki et al., 2006; van Leeuwen and Bouten, 2004; van Leeuwen et al., 2006; Yeung et al., 2004].

It is important to note that, unless some explicit cases, we do not consider other performance indicators strongly related to particular implementations and hardware/software architectures. These indicators include the computational efficiency, and the computational resources used in terms of storage and memory allocation [Cappelli et al., 2006]. In this Thesis, basic implementations of the strategies studied have been tested on diverse configurations of computational resources, all of them running Debian Linux [Debian].

3.1.1. Statistical significance of experimental results

Guyon et al. [1998] derived the minimum size of the test data set, \( N \), that guarantees
statistical significance in a pattern recognition task. The goal was to estimate $N$ so that it is guaranteed, with a risk $\alpha$ of being wrong, that the probability of error $P$ does not exceed the estimated error rate from the test set. As an example, for a typical configuration (0.05 and 0.2, respectively), the following simplified criterion is obtained:

$$N \approx \frac{100}{P}$$  \hspace{1cm} (3.1)

There are also simplified mnemonic rules in order to establish an approximate value of $N$ which satisfies the statistical significance of estimated error rates. One of the most popular in the automatic speaker recognition community is the so-called Doddington’s rule of 30 [Doddington, 1998], which is stated as follows:

*To be 90% confident that you are within 30% of the true error rate, you need at least 30 errors.*

The experimental protocols in NIST SRE from 2004, as they will be used in this Thesis, have been designed to be statistically significant, as it is explained by [van Leeuwen et al. 2006].

### 3.2. Experimental framework

This section presents the databases, protocols and baseline systems used in this Thesis in order to present experimental results.

#### 3.2.1. Databases

One key element for performance evaluation of speaker recognition systems is the availability of speech databases. The representativeness of speech features corresponding to a large population of individuals, together with the desirable presence of all possible factors of variability of the speech signal (i.e., multi-session, multiple languages, multiple environmental conditions, etc.), makes database collection a time-consuming and complicated process, in which a high degree of co-operation of the donors is needed. Additionally, the legal issues regarding data protection are controversial [Wayman et al., 2005b]. However, the efforts of institutions [ELRA; LDC; NIST] and projects [Campbell and Higgins; CAVE; Garofolo et al.; SPEECHDAT] regarding speech database collection and benchmarks using common corpora have led to a significant amount of publicly available databases for speaker verification.

In this section we describe the public databases to be used in this Thesis, as well as other existing corpora used in automatic speaker recognition. As the experimental work presented in this Thesis has been mainly based on the NIST SRE protocol from 2004 to the last 2006 edition, we will concentrate on the databases used in such campaigns.

#### 3.2.1.1. NIST SRE databases

Table 3.1 summarizes the corpora used in NIST evaluations. Other information is also summarized in the table, such as language variability; environmental variability (channel, network,
3. EXPERIMENTAL FRAMEWORK

Table 3.1: Databases and conditions of NIST SRE corpora. Source: van Leeuwen et al. [2006].

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sw2p3</td>
<td>sw2</td>
<td>sw3p1</td>
<td>sw3p2</td>
<td>sw3p2</td>
<td>MIXER</td>
<td>MIXER</td>
<td>MIXER</td>
</tr>
<tr>
<td>Ahumada</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Languages</td>
<td>English</td>
<td>English/</td>
<td>English</td>
<td>English</td>
<td>English/</td>
<td>English/</td>
<td>English/</td>
<td>English/</td>
</tr>
<tr>
<td># Speakers</td>
<td>233</td>
<td>804</td>
<td>174</td>
<td>330</td>
<td>356</td>
<td>310</td>
<td>526</td>
<td>1011</td>
</tr>
<tr>
<td>Channel</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
</tr>
<tr>
<td></td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
</tr>
<tr>
<td></td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
<td>Land</td>
</tr>
<tr>
<td>Train length</td>
<td>2m</td>
<td>2m</td>
<td>2m</td>
<td>2m</td>
<td>2m</td>
<td>2m</td>
<td>2m</td>
<td>2m</td>
</tr>
<tr>
<td>Conv. sides</td>
<td>1,2,4,8,16</td>
<td>4,8,16</td>
<td>1,3,8,16</td>
<td>1,3,8,16</td>
<td>1,3,8,16</td>
<td>1,3,8,16</td>
<td>1,3,8,16</td>
<td>1,3,8,16</td>
</tr>
<tr>
<td>Test length</td>
<td>3−60s</td>
<td>5−60s</td>
<td>5−60s</td>
<td>5−60s</td>
<td>5−60s</td>
<td>10,30s</td>
<td>10,30s</td>
<td>10,30s</td>
</tr>
<tr>
<td>Conv. sides</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Switchboard 1 (sw1).** It contains conversational speech in American English, recorded over landline telephone. Dialectal variability is not considered. It contains speech recorded from different telephone lines and from carbon-button and electret telephone handsets. These sources of variability have demonstrated to seriously affect recognition performance [Reynolds et al., 2000].

**Switchboard 2 (sw2).** It contains conversational speech in American English recorded over landline telephone. As Switchboard I, it contains different lines and handset variability, but this time in a higher degree. This database was recorded in three phases containing different dialectal content: Phase 1 (sw2p1, mid-Atlantic US English), Phase 2 (sw2p2, mid-Western US English) and Phase 3 (sw2p3, Southern US English).

**Switchboard 3 (sw3), also known as Switchboard Cellular.** It contains conversational speech in American English recorded over cellular networks. It has been recorded in two phases containing different channel content: Phase 1 (sw3p1, GSM transmission channel) and Phase 2 (sw3p2, CDMA transmission channel). It is recorded in American English (several dialects).

**MIXER and additional multi-language data.** MIXER presents three fundamental differences with respect to all versions of Switchboard. First, the channel and handset variability is significantly higher, including speech recorded over cordless telephones in landline and cellular networks and handsets such as ear buds, head-mounted phones and speaker phones, as well as regular handsets. Second, it is multilingual, containing speech in American English, Spanish, Arabic, Mandarin Chinese and Russian. Third, a novel Fisher protocol was
used in order to randomize conversations between donors in the database. Thus, MIXER contains much more variability than previous databases. Moreover, for NIST SRE 2005 and 2006 campaigns a significant amount of new speech data from the former languages was recorded following the same protocol, including dialect variation and non-native speakers. Details about the MIXER database can be found in [Campbell et al. 2004a; Przybocki et al. 2007].

**Multi-microphone data.** During NIST SRE 2005 and 2006 campaigns, a significant effort was devoted to collect multi-microphone databases. The idea is acquiring a conversation from a speaker simultaneously through landline telephone and a variety of microphones, namely: an ear-bud/lapel mike, a mini-boom mike, a courtroom mike, a conference room mike, a distant mike, a near-field mike, a PC stand mike and a micro-cassette mike. On the one hand, this acquisition protocol is obviously much more complicated, as the speech donors have to be present where the acquisition hardware is installed (telephone and multiple microphones). On the other hand, the variability of the conditions is much higher than for telephone speech, as the position of the speaker and the background noise and reverberation severely affect the response of the microphones. This so-called multi-channel, auxiliary-microphone or cross-channel framework constitutes a challenging and realistic benchmark which is recently attracting significant research [Sturim et al., 2007]. For this reason, the main task in NIST SRE 2008 is planned to be mainly based on auxiliary-microphone data [NIST].

**Ahumada.** The database Ahumada was recorded by the ATVS group. It contains speech in Spanish recorded over the telephone and two types of microphones under controlled conditions. There is variability in the speaking style, from read text to spontaneous speech. Details about the Ahumada corpus can be found in [Ortega-Garcia et al. 2000].

**FBI database.** The FBI database was proposed for NIST SRE 2002 in order to model the effect of speech data with different microphone channels. This is a typical situation in forensic automatic speaker recognition, where the questioned and incriminating speech material may be a wire-tapped telephonic conversation and the control material is acquired with different microphones in police dependencies. The development of such database has encouraged the acquisition of speech data coming from different microphones, whose importance has increased in recent NIST SRE. Details about this multichannel database can be found in [Nakasone and Beck 2001].

Other databases for automatic speaker recognition not included in NIST SRE corpora are the TIMIT database [Garofolo et al.], the YOHO database [Campbell and Higgins] and the BioSec database [Fierrez et al., 2007]. The two last ones are mainly used for text-dependent applications.
3. EXPERIMENTAL FRAMEWORK

3.2.2. Experimental protocols

In this Thesis we will follow the protocols defined in NIST SRE 2005 and 2006. The basic task in NIST SRE is the detection or verification task. For each training speech segment and test segment, the system must output a decision of whether the two speech fragments belong to the same person or not.

A training condition is defined by NIST as the length of the speech segment used for training on each detection trial, and similarly the test segment condition is defined for the test segment. The NIST evaluation protocol defines since 2004 the following training conditions: 10 seconds, 1, 3, 8 and 16 conversation sides. Also, the following test segment conditions are defined: 10 seconds, 1 conversation side, 3 full conversations in a mixed channel and auxiliary microphone data. Each conversation side is also known as a 4-wire (4w) segment, and two conversations mixed in a single channel is also known as a 2-wire (2w) segment. Each conversation side has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Although there are speakers of both genders in the corpus, no cross-gender trials are defined.

The combination of a training and a test segment condition defines a test condition, which constitutes the minimum test that a participant can perform in a NIST SRE. A test condition consists of a given set of detection trials with common training and test segment conditions. One among the possible test condition is always designated as the core or required test condition, being mandatory to all participants. In NIST 2004, 2005 and 2006 campaigns the core test condition has been the combination among 1 conversation side for training and 1 conversation side for testing. In this Thesis, this condition will be referred to as 1conv4w-1conv4w, 1c-1c, 1side-1side or 1s-1s.

Not all the test conditions are allowed in a given NIST SRE. In Table 3.2, a summary of the conditions in each evaluation since 2004 is represented. The core condition (1conv4w-1conv4w) appears in boldface for each year.

The test conditions used in this Thesis correspond to the 1conv4w-1conv4w (1c-1c) and 8conv4w-1conv4w (8c-1c) conditions of NIST SRE 2005 and 2006 campaigns, as it is shown in Table 3.2. The number of target and non-target trials (scores) for each of those test conditions is shown in Table 3.3. It can be seen that for all cases the statistical significance is satisfactory according to van Leeuwen et al. [2006].

Details about the NIST SRE 2005 and 2006 protocols can be found in the corresponding evaluation plans NIST [2005, 2006], in the NIST webpage NIST and in Przybocki et al. [2007]; van Leeuwen and Bouten [2004].

Apart from the evaluation protocol, a database labeled with the true speaker identities is needed in order to generate background models, populations, normalization cohorts, fusion training scores, etc. This database is referred to as the development database in NIST SRE. Generally, all the available databases for each of the participants are allowed to be used as development data, except the evaluation database used in the given NIST SRE. Eventually, a development set is available for participants through the LDC LDC, containing all the devel-
### Table 3.2: NIST SRE test conditions since SRE 2004. The core condition for each year appears in boldface. Test conditions used in this Thesis appear in italic (1conv. 4w-1conv. 4w and 8conv. 4w-1conv. 4w in both SRE 2005 and 2006).
### Table 3.3: Sample size of NIST SRE test conditions used in this Thesis.

<table>
<thead>
<tr>
<th>Condition</th>
<th># Target Scores</th>
<th># Non-target Scores</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1c-1c NIST SRE 2005</td>
<td>2728</td>
<td>28033</td>
<td>30761</td>
</tr>
<tr>
<td>1c-1c NIST SRE 2006</td>
<td>3612</td>
<td>47836</td>
<td>51448</td>
</tr>
<tr>
<td>8c-1c NIST SRE 2005</td>
<td>2246</td>
<td>21426</td>
<td>23672</td>
</tr>
<tr>
<td>8c-1c NIST SRE 2006</td>
<td>2956</td>
<td>29553</td>
<td>32509</td>
</tr>
</tbody>
</table>

development and evaluation databases of past NIST SRE. However, the use of the development data is dependent on the experiment, and not restricted by NIST\(^1\). Thus, it will be detailed in each particular experimental section.

#### 3.2.3. Baseline automatic speaker recognition systems

In this section we describe the score-based baseline automatic speaker recognition systems used in this Thesis. It is important to highlight that the presented baseline systems generate a score for each trial, or in other words, no LR computation technique is used in the baseline system configuration. However, these systems will be used in order to generate the scores needed for the LR computation techniques used in this Thesis.

All the systems presented in this Section have been developed at the ATVS-UAM Biometric Recognition Group [ATVS].

**GMM system description.** The ATVS-UAM GMM system is based on a detector with target and alternative probability distributions modelled by Gaussian Mixture Models [Reynolds et al., 2000]. The similarity scores are computed by means of the following formula:

\[
s(O, \lambda_t) = \log p(O | \lambda_t) - \log p(O | \lambda_{UBM})
\]  

where \(p(O | \lambda_t)\) and \(p(O | \lambda_{UBM})\) are the probability density functions for the target model and a Universal Background Model (UBM), which are modeled as mixtures of Gaussians [Reynolds et al., 2000], and \(O\) is the sequence of feature vectors from the speech test utterance. Target models are derived using MAP adaptation from the UBM using the Expectation Maximization (EM) algorithm [Duda et al., 2001].

Feature extraction in order to obtain the \(O\) sequence for each utterance is performed as follows: 19 Mel Frequency Cepstral Coefficients (MFCC) are extracted from the speech signal using overlapped Hamming windows [Deller et al., 1999]. RASTA filtering [Hernansky and Morgan, 1994] is further applied to the parameters. Feature warping [Pelecanos and Sridharan, 2001] was firstly used in order to compensate channel effects. However,

\(^1\)Unless, of course, the prohibition of not using the same evaluation data as development
it was replaced by feature mapping in NIST SRE 2006 [Reynolds, 2003a]. The channel compensation technique will be specified in each of the presented experiments.

Finally, T-Norm score normalization has been used in order to normalize effects in the score distributions due to variabilities in the speech test segment [Auckenthaler et al., 2000].

**SVM-GLDS system description.** The ATVS-UAM SVM-GLDS speaker recognition system is also based on the spectral characteristics of the speech, as the GMM system described above. However, in this case the similarity score computation is based on discriminative Support Vector Machines (SVM) [Campbell et al., 2006a]. A kernel expansion is performed on the whole observation sequence $O$, and a separating hyperplane is computed between the speaker features and the background model.

The ATVS-UAM SVM-GLDS system uses a polynomial kernel of degree 2 or 3 (will be specified in the experimental framework), showing a significant improvement from degree 2 to degree 3 [Wan and Campbell, 2000]. The expansion is followed by a Generalized Linear Discriminant Sequence Kernel (GLDS) as described in Campbell [2002]. We use a channel compensation matrix [Solomonoff et al., 2004] in order to avoid channel mismatch effects.

The system uses the same feature extractor as the GMM system. Two gender-dependent and channel-independent data sets from the development set are used for background modeling. T-Norm has also been used for score normalization.

**GMM-SVM-SV system description.** The ATVS-UAM GMM-SVM-SV system is based on Support Vector Machine (SVM) classification of the means of GMM models trained from the speech features [Campbell et al., 2006b].

For each speech utterance, a GMM model is generated using the ATVS-UAM GMM system described above. The parameters of a GMM which is adapted using only-means-MAP from the same UBM differ only in the mean vectors. Therefore, each GMM is represented as a supervector in a high-dimensional space, constructed by concatenating the mean vectors of each mixture.

In order to compensate for intersession problems, a technique known as Nuisance Attribute Projection (NAP) has been applied [Solomonoff et al., 2005]. This technique projects each supervector to a subspace where the variability due to channel conditions is minimized. T-Norm is also performed for score normalization.

### 3.3. Chapter summary and conclusions

In this chapter we have described the experimental framework used for presenting results through all this Dissertation. The experimental framework has been organized as follows. First, the evaluation protocol followed in the Thesis was introduced. Second, databases used for generating results were described, as well as other corpora used in automatic speaker recognition.
Third, NIST SRE protocol was described as a public benchmark in order to compare results. Finally, several automatic speaker recognition systems developed by ATVS-UAM (Biometric Recognition Group) and used in this Thesis have been described.
Chapter 4

A hierarchical methodology for forensic automatic speaker recognition

This chapter proposes a hierarchical methodology for the use of automatic speaker recognition systems in forensic identification. Although forensic automatic speaker recognition is nowadays a mature technology useful in real applications, an essential attention has to be taken to the upcoming requirements of forensic science, which affect the theoretical and practical procedures and methodologies of any forensic discipline. Driving forces of this fundamental change are the admissibility of the forensic disciplines in court and the scientific foundations of the methods. Therefore, this methodology supposes a powerful tool for practitioners, as it allows transparent and testable forensic analysis of the evidence using the common score-based architecture in most automatic speaker recognition systems.

According to such requirements, and taking into account the state of the art in automatic speaker recognition technologies, in this chapter we identify a hierarchy of abstraction levels in the forensic automatic speaker recognition process: the discrimination level, the presentation level and the forensic level. Each level has a different function, aiming at presenting evidence evaluation in court in a transparent and scientific way.

This chapter is organized as follows. First, we describe the motivation of what has been dubbed the coming paradigm shift in forensic science [Saks and Koehler, 2005], identifying its main requirements. Then, we establish the objectives of our methodology according to such requirements. The methodology is then presented according to a hierarchy of abstraction levels in the forensic automatic speaker recognition process. Finally, conclusions are drawn.

The contribution of this chapter is summarized as follows: i) the identification of the requirements from the coming paradigm shift which affects forensic automatic speaker recognition; ii) the identification of the common methodologies in automatic speaker recognition with respect to forensic identification; iii) the proposal of a methodology for forensic automatic speaker recog-
4. A HIERARCHICAL METHODOLOGY FOR FORENSIC AUTOMATIC SPEAKER RECOGNITION

nition systems according to the identified needs; and iv) the definition of the proposed hierarchy of abstraction levels in the evidence evaluation, interpretation and reporting process.

4.1. The coming paradigm shift in forensic identification science

In Chapter 1 we introduced the debate currently being held about the scientific foundations of many disciplines in forensic science. One of the main streams of such debate advocates for what has been dubbed the coming paradigm shift in forensic identification science. As cited there, the term should not be understood as a Kuhnian paradigm shift [Kuhn, 1962], but as a metaphor highlighting the transformation of forensic science from a pre-science into a well-founded scientific discipline. According to Saks and Koehler [2005], there are four main driving forces motivating this substantial change:

Changes in the law. The Daubert rules [U.S. Supreme Court, 1993] have established a first step on the requirements of admissibility in the USA. Briefly, in order to be admitted in court, any technique must satisfy the following conditions: i) it has been or can be tested, ii) it has been subjected to peer review or publication, iii) there exist standards controlling its use, iv) it is generally accepted in the scientific community, and v) it has a known or potential (and acceptable) error rate. As highlighted by Saks and Koehler [2005], the Daubert rules lower the admissibility threshold for scientifically sound techniques based on transparent forensic testing and raises the threshold for methodologies lacking of a scientific foundation, such as experience-based testimony. Thus, the Daubert rules scrutinize forensic disciplines, and constitute a difficult handicap for techniques lacking of supporting data or reporting obscure and/or non-repeatable conclusions. Hence, Daubert rules give judges the gatekeeper role in order to admit scientific evidence. Although Daubert rules belong to the US jurisprudence, their impact on the community all over the world has been remarkable, as they establish a severe conceptual change in practitioners and legal parties, which is in accordance to the opinions of many experts worldwide [Champod, 2006; Champod and Evett, 2001; Cole, 2005a; Kennedy, 2003; Saks and Koehler, 2005].

Wrongful convictions. The evidence of errors in identifications has sounded the alarm about the assumed infallibility of some forensic disciplines [Scheck et al., 2000]. A remarkable example, highlighted by the mass media, is the erroneous fingerprint identification in the Mayfield case in March 2004 Madrid bombings [DoJ, 2006]. In many cases, such erroneous identifications have led to convictions. This fact has motivated initiatives like the Innocence Project, where DNA analysis has been applied on crime scene evidence for which no test was performed at the time of the trial. That has led to the exoneration of a significant number of individuals previously convicted in the past (about 200 by 2007). In Saks and Koehler [2005], 85 of these exoneration cases are analyzed looking for the reasons of such wrongful convictions, and the results are shown in Figure 4.1. It is shown that forensic science testing errors is one of the main causes associated to such wrongful
4.1 The coming paradigm shift in forensic identification science

convictions. Moreover, false or misleading testimony by forensic scientists is also present in a significant amount of cases. Other reported cases and studies may be found in Cole [2005]; Drod [2006]; Scheck et al. [2000].

Computation of error rates. Although it is a clear Daubert requirement, performance testing is not common in forensic science. The lack of research about data-driven experimental assessment in many forensic disciplines does not help to scientifically state the accuracy of the methods. Moreover, in Fagman et al. [2002]; Saks and Koehler [2005] some performance measures (mainly error rates) have been computed from forensic tests performed in disciplines like spectrographic voice analysis, handwriting, bite marks and mitochondrial DNA profiling, yielding surprisingly poor results in some cases. It is remarked that, in order to foster scientific testing of techniques, data collection and availability is a key issue.

DNA as a scientific model. In Saks and Koehler [2005], DNA analysis is addressed as a model of scientific forensic discipline. There are three main reasons supporting such appointment. First, since its origin in the 1980s DNA profiling has been scientifically based, which has favored avoiding expert opinions based on experience. Second, its transparent and clear procedures have facilitated scrutinizing and inspection by fact finders in individual cases, and therefore non-scientific methods have been early eliminated. Third, DNA profiling is data-driven and probabilistic, avoiding hard match or non-match statements, and grounded on an experimental and repeatable framework supported by large databases.

Figure 4.1: Factors present in 85 wrongful convictions, based on the case analysis data from the Innocence Project. The two highlighted bars indicate causes related to forensic science. Source: Saks and Koehler [2005].
of DNA profiles. Also, probabilistic opinions in DNA can easily be expressed as \( LR \) values in a Bayesian context, as a number of experts advocate for [Aitken and Taron, 2004; Champod and Meuwly, 2000; Evett, 1991; Evett and Buckleton, 1996].

From the above driving forces as described in Saks and Koehler [2005], several immediate needs can be derived for any forensic discipline to follow the coming paradigm shift.

1. **Transparency of procedures.** This is essential to allow fact finders and scientists to continuously scrutinize the methods and to identify and eliminate non-scientific practices. Also, clarity in the presentation of forensic results will help fact finders to correctly interpret the weight of the scientific evidence, as well as the accuracy of the forensic discipline at hand. Transparent reporting of the databases used for evidence evaluation is also necessary.

2. **Testability of the techniques.** The determination of the accuracy of a scientific forensic discipline should be based on experimental results from data-driven tests, representing real-case conditions if possible. In this sense, data collection and availability is fundamental in order to perform repeatable experiments. Also, common forensic testing methodologies would help in sharing resources from different institutions, aiming at comparing and improving the performance of different techniques.

3. **Accuracy.** Accuracy can be defined as the degree of conformity of a measured or calculated quantity to its actual (true) value. In forensic identification at a source level [Cook et al., 1998] the quantity to calculate is which one of the competing hypothesis is true, e.g. whether the control and recovered speech material come from the same source or not. Daubert rules state that forensic disciplines should have been tested, and their accuracy obtained, before determining their admissibility. Hence, it is important for the accuracy of techniques to be continuously improved in order to face real-world challenges. Furthermore, it is essential to define a proper and common selection of performance measures clear for scientists and fact finders, in order to avoid confusion or misleading practices when stating the accuracy of forensic disciplines.

4. **Common procedures.** It is important that practitioners adopt common methodologies to present results in court. This is required not only with regards to the admissibility requirements of the Daubert rules, but also in order to avoid wrong or misleading conclusions due to a disparity of practices among (and even within) disciplines. This convergence is desirable at any stage in the forensic identification process, from the evidence evaluation and interpretation methodology to the testing and experimental framework which determines the accuracy of the discipline at hand.

It seems clear from the previous descriptions that the identified requirements are interrelated among them, as it is shown in Figure 4.2.
4.2. Forensic automatic speaker recognition in the steps of DNA

DNA analysis accomplishes to a large extent the requirements of the coming paradigm shift, identified in the last section. First, DNA is transparent, with a clear probabilistic methodology, often based on $LR$ analysis of the evidence. Transparency is also present in the reporting of results from evidence evaluation with the use of clearly defined databases. Second, DNA is testable, as repeatable data-driven experimental procedures for testing allows the scientific statement of the accuracy of the techniques. Third, DNA is accurate, having high performance for many typical measures in forensic science. Finally, due to the efforts of forensic laboratories, government agents and legal entities, DNA has similar procedures in many places in the world, allowing inter-operability, rapid coordination between institutions and a common language in order to understand results from forensic reports [Butler, 2005; Council, 1992, 1996].

In order to fulfil the proposed needs for forensic automatic speaker recognition, the methodology proposed in this Thesis seeks emulating DNA in its procedures and protocols for evidence evaluation and forensic testing [Gonzalez-Rodriguez et al., 2007b], as it is described below:

1. **Transparency.** The two main desiderata for a transparent framework are: i) that it is based on scientific principles, and ii) that the analysis procedures and reporting of results perfectly accord with the needs and roles of both forensic scientist and other interested legal parties. On the one hand, automatic speaker recognition is based on signal processing and pattern recognition techniques. Both of them are widely accepted scientific disciplines. On the other hand, the use of a probabilistic framework as proposed in [Saks and Koehler, 2005] allows the reporting of meaningful values to the court based on the analysis of the objective data in the case, avoiding non-scientific statements such as hard identification/exclusion conclusions [Champod, 2000]. Moreover, a $LR$ framework for the analysis of the evidence for forensic automatic speaker recognition as proposed in the literature [Aitken and Taroni, 2004; Evett, 1998] allows to clearly determine the roles of forensic science practitioners and fact finders [Champod and Meuwly, 2000]. With such a framework, the whole evidence evaluation process can be understood and scrutinized by both fact finders and forensic
2. **Testability.** In order to know the performance of a forensic technique in a data-driven and repeatable way, speech database collection and unified forensic testing protocols are needed. In this sense, the work by NIST and NFI-TNO in their respective SRE [Przybocki *et al.*, 2007; van Leeuwen *et al.*, 2006] has been fundamental for automatic speaker recognition in the last years. A known protocol and a well documented database allows the accuracy of a forensic automatic speaker recognition technique to be determined in controlled and transparent conditions for the court.

3. **Accuracy.** Daubert rules state that the accuracy of forensic disciplines should be known in order to determine their admissibility. Although several performance figures are referred to in the literature, as a measure of accuracy, it is important to clearly select such a measure (e.g., *error rates* are often referred to in Saks and Koehler [2005]). Furthermore, if *LR* methodology for evidence evaluation is used, there is a need of addressing proper performance measures beyond *error rates*. Moreover the accuracy of the technique in use is important for its admissibility in courts, and it is no secret that disciplines presenting high accuracy such as fingerprints and DNA are more popular in trials than disciplines presenting a lower accuracy. Forensic automatic speaker recognition has to take this into account in order to face real-world challenges [Bonastre *et al.*, 2003].

4. **Common procedures.** The use of common procedures in order to report results in forensic automatic speaker recognition is a key issue for admissibility, as it is stated by the Daubert rules. This may be applied to all of the identified needs of the *coming paradigm shift* with regards to automatic speaker recognition systems, as have been discussed above.

In Figure 4.3 the proposed key lines for adapting forensic automatic speaker recognition to the requirements of the *coming paradigm shift* in forensic identification are summarized.

### 4.3. Common procedures in automatic speaker recognition

In order to accomplish the lines proposed above for forensic automatic speaker recognition, it is needed to consider the accepted procedures in the state of the art in automatic speaker recognition. Some main established commonalities can be identified as follows:

- **Score-based architecture.** As it was described in Chapter 2 the vast majority of automatic speaker recognition systems generate a score as a measure of similarity between two sets of speech material, with independence of the identity levels exploited, the feature extraction scheme, the modeling strategy, the fusion method, etc.

- **Existence of widespread benchmarks.** NIST and NFI-TNO speaker recognition evaluations [Przybocki *et al.*, 2007; van Leeuwen *et al.*, 2006] have fostered research during the last years, and have become *de facto* standards in the assessment of automatic speaker recognition.
4.3 Common procedures in automatic speaker recognition

**Transparency**
- Based on scientific, accepted disciplines (pattern recognition and signal processing)
- LR probabilistic framework (DNA)

**Testability**
- Collection and clear definition (and public availability if possible) of databases
- Clear and defined forensic testing protocols

**Accuracy**
- Proper assessment in a LR framework
- Clear presentation in court
- Constant improvement

**Common procedures**
- For transparency
- For testability
- For accuracy

Figure 4.3: Identified lines to follow in the proposed methodology in order to adapt forensic automatic speaker recognition to the coming paradigm shift in forensic science.

Recognition technology. They have provided testing protocols, challenging and publicly available databases and assessment methodologies. The proposed framework for forensic automatic speaker recognition should consider such procedures in testing protocols and performance evaluation measures.

- **Emphasis on discriminating power.** During the last decade, the main measure of performance found in the literature for automatic speaker recognition systems, as well as most influential benchmarks, has been based on discriminating power using e.g. DET plots [Martin et al., 1997]. Although it is not the only metric to take into account in a LR framework [Dessimoz and Champod, 2007], it is desirable that LR computation will preserve the discriminating power of scores yielded by an automatic speaker recognition system.

- **Independence of prior information of the score.** As it will be shown in Chapter 6 for a detection task, the prior probabilities about the relevant hypotheses must be given and it is not the role of the system to compute them [Brümmer and du Preez, 2006]. This is in accordance with the use of prior probabilities and decision costs in most popular benchmarks in order to compute measures of performance based on mean cost [Przybocki et al., 2007]. This approach agrees with the Bayesian framework for evidence evaluation and interpretation, and the proposed methodology should take advantage of it in order to be compliant to the established procedures.

---

1 DET plots will be described in detail in Chapter 5.
4. Hierarchical methodology for forensic automatic speaker recognition

The methodology proposed in this section considers the lines identified as a result of the coming paradigm shift and the common procedures accepted by the automatic speaker recognition community in order to define a hierarchical methodology for forensic automatic speaker recognition (Figure 4.4). The methodology can be stated in the following steps:

1. **Computation of the similarity score** between the control and recovered speech materials using an automatic speaker recognition system.

2. **Computation of a LR value.** At this stage, the score generated by the automatic speaker recognition system is transformed into a LR value.

3. **Reporting of results to the court.** The reported LR value measures the weight of the forensic evidence in the inferential process for the given case, as it happens in DNA analysis. Also, a proper measure of the accuracy of the technique should be included in the report in case the fact finder will question the admissibility of the procedure, or according to court demands.

4.4.1. Hierarchy of abstraction levels

According to the methodology proposed above, we clearly identify three abstraction levels in the forensic speaker recognition process:

1. **The discrimination level: score-based automatic speaker recognition.** The objective of this level is yielding a similarity score between the recovered and control speech materials. The inputs of this level are the recovered and control speech material, as well as speech from databases in order to model background knowledge or normalize the output score. The objective at this level is achieving the best possible discriminating power, as this has been one of the main objectives in automatic speaker recognition in the last years [Bimbot et al., 2004; Reynolds, 2003b; van Leeuwen et al., 2006]. The output of this level will be the similarity score, as the wide majority of automatic speaker recognition systems are score-based. The results of testing procedures aimed at measuring the discriminating power of the score are also included as outputs of this level. This abstraction level allows the system to work in detection (verification) or identification modes, and therefore in many commercial applications of automatic speaker recognition technology (see Chapter 2).

2. **The presentation level: LR computation.** The objective of this level is transforming the score yielded by the discrimination level into a LR value, which allows the obtention of posterior probabilities in a Bayesian framework. The inputs of this level are the score from the discrimination level and speech databases, which will be necessary in order to obtain
4.4 Hierarchical methodology for forensic automatic speaker recognition

Proposed methodology for forensic speaker recognition

- Transparency
- Testability
- Accuracy
- Common procedures

Score-based
- Benchmarks
- Discriminating power
- Priors & costs independent of the score

Figure 4.4: Elements considered by the proposed methodology for forensic automatic speaker recognition.

Figure 4.5: Abstraction levels in the proposed methodology for forensic automatic speaker recognition.
the objective score-to-$LR$ transformation. The output of this level will be the $LR$ value, which could be used in order to obtain posterior probabilities; and performance measures aimed at measuring the accuracy of the computed $LR$ value (both the discriminating power and the calibration loss).

3. **The forensic level: forensic reporting.** The objective of this abstraction level is reporting the results of forensic analysis to a court of law according to the requirements addressed above (Figure 4.3), by means of the statement of the weight of the evidence and the accuracy of the technique in a transparent, scientific and repeatable way. The inputs of the forensic level are the $LR$ value from the presentation level and the information about databases used in its computation. On the other hand, court requirements are also taken into account at this level, and speech databases should be also necessary in order to assess the technique in a case-adapted scenario. The outputs of this level are the forensic report, which should include the $LR$ value and should be also adapted to the court needs, and assessment results in order to show the accuracy of the whole process in a transparent way.

This classification will structure the following chapters in this Thesis, and is illustrated in Figure 4.5.

4.5. **Chapter summary and conclusions**

In this chapter we have defined a hierarchical methodology for forensic automatic speaker recognition. First of all, we have analyzed the situation of forensic science in recent years, particularly focusing on the current debate about admissibility and scientific procedures used in forensic disciplines. We have deeply described which has been dubbed as the *coming paradigm shift* in forensic identification science [Saks and Koehler, 2005], which claims for a change to more scientific procedures in many classical forensic identification techniques.

On the one hand, we have derived some main requirements which a forensic discipline should satisfy for being considered scientific and acceptable in court, addressing lines of action based on DNA profiling in order to fulfil them. On the other hand, common and widely accepted procedures in automatic speaker recognition have been identified from the state of the art. The proposed methodology has been then stated as a way of using automatic speaker recognition systems for forensic identification, considering both the *coming paradigm shift* in forensic identification and the commonly accepted methods in automatic speaker recognition. Therefore, the proposed methodology allows the use of typical score-based automatic speaker recognition systems for scientific forensic identification without specific changes in its architecture.

The analysis on the requirements demanded in forensic science and automatic speaker recognition described in this chapter are original contributions.
Chapter 5

The discrimination level: score-based automatic speaker recognition

This chapter presents the discrimination level of the proposed hierarchical methodology for forensic automatic speaker recognition. The objective of this level is computing a similarity score between the recovered and control speech materials. Such a score has been the typical output of automatic speaker recognition systems in the last decade, present in the majority of text-independent automatic speaker recognition systems [Bimbot et al., 2004]. Moreover, score-based automatic detection and identification systems have shown their adequacy in many applications such as speaker authentication, and speaker spotting. The discrimination level belongs to the proposed hierarchical methodology for forensic automatic speaker recognition, and it has been defined in Chapter 4. In Figure 5.1 the inputs and outputs of this level are detailed. On the one hand, this level receives the following inputs:

- **Recovered and control speech material.** The comparison of both elements by the automatic speaker recognition system will generate a similarity score.

- **Speech databases**, in order to build background models or score normalization cohorts.

- **Background information about the case.** The score computation process will benefit from the knowledge about the circumstances and conditions of the recovered and control speech segments. This will allow a proper selection of background data and normalization cohorts in order to improve the discriminating power of the automatic speaker recognition system.

On the other hand, the discrimination level yields the following outputs, which will be inputs to the following level (the presentation level, described in Chapter 6):

- **Similarity score.** The higher the support to the same-source hypothesis for the control and speech material, the higher the score, and vice-versa. This value will allow next
levels in the hierarchy to evaluate the evidential weight of the control-recovered speech comparison.

- **Testing results.** Results of testing procedures aimed at measuring the discriminating power of the score are also included as outputs of this level. They will allow next levels to assess the variation of discriminating power due to the score-to-LR transformation.

This chapter is organized as follows. First, we discuss the concept of discriminating power, concluding with a definition which will be appropriate considering the automatic speaker recognition state of the art, to which we will adhere in this Thesis. Performance measures of discriminating power found in the literature will be also presented. Then, experimental results will be reported in order to illustrate several techniques commonly used for improving the discriminating power in automatic speaker recognition: score normalization, session variability compensation and fusion of systems. A contribution of this Thesis on fast score normalization for the improvement of discriminating power, namely KL-T-Norm, is also fully addressed, showing experimental results which demonstrate its adequacy. Finally, conclusions are drawn.

The contribution of this chapter relies on the proposed KL-T-Norm score normalization technique. All experiments in this chapter have been performed using the NIST SRE 2005 and 2006 protocols and databases, described in Chapter 3.

### 5.1. Discriminating power

The question of how good is a technique in order to distinguish between two samples of material has been a matter of discussion since the middle of the 20th century [Aitken and Taron].
5.1 Discriminating power

![Graphs showing False Acceptance (FA) and False Rejection (FR) rates for two sets of scores presenting perfect (a) and non-perfect (b) discriminating power, with respect to the decision threshold \( \tau \).](image)

Figure 5.2: FA and FR rates curves for two sets of scores presenting perfect (a) and non-perfect (b) discriminating power, with respect to the decision threshold \( \tau \).

In order to address this problem, discriminating power has been defined for discrete data in the context of DNA population genetics as the probability of two samples extracted from a population will match if they come from different sources. In Aitken and Taroni [2004] a detailed review about the use of discriminating power for discrete data can be found.

In automatic speaker recognition, discriminating power is associated with correctly discriminating same-source and different-source trials. In fact, discriminating power (also known as discrimination) has been the main performance measure for automatic speaker recognition systems in the last years [Przybocki et al., 2007; van Leeuwen et al., 2006].

5.1.1. Discriminating power of a set of scores

Discriminating power will be defined here in a data-driven way for a set of same-source (target) and different-source (non-target) scores. Consider an assessment test conducted with a database of speech samples where the sources are known. Target and non-target scores are then generated by respectively comparing same-source and different-source speech samples with an automatic speaker recognition system.

In a speaker detection system, a decision should be taken, and therefore a binary classification problem is defined. The two classes of the binary classification problem correspond to a target (same-source) non-target (different-source) trials. In order to take such decision, a given decision threshold \( \tau \) is compared to a similarity score, leading to a decision that the given trial is a target trial or a non-target trial. These decisions are respectively referred to acceptance and rejection decisions in the literature [Bimbot et al., 2004], in relation to access control applications.

Such a binary decision framework involves a tradeoff between two types of errors: 1) False Rejection error (FR), occurring when it is decided that the control and recovered speech material come from different sources, when they actually come from the same source; and 2) False
Acceptance error (FA), taking place when it is decided that the control and recovered speech material come from the same sources when they actually come from different sources. Although each type of error can be computed for a given decision threshold $\tau$, a single performance level is inadequate to represent the full capabilities of the system. Therefore the performance capabilities of detection systems have been traditionally shown in the form of FA and FR Rates versus the decision threshold $\tau$, as depicted in Fig. 5.2 for an ideal system (a), and a typical system (b).

FA and FR rates in the score set being evaluated will be different if the value of $\tau$ changes. The discriminating power, or discrimination performance of a set of scores is defined as follows. Two different sets of scores have the same discrimination performance if, for every possible threshold $\tau$ in the first score set, a threshold $\tau'$ can be found for the second score set such as the false acceptance and false rejection rates are the same for both score sets. Conversely, for every possible threshold in the second set $\tau'$, there exist a threshold in the first set $\tau$ leading to the same FA and FR rates.

False acceptance and false rejection probabilities can be assigned to these threshold-dependent rates, leading to:

- $P_{fa}(\tau)$, the probability of a false acceptance.
- $P_{fr}(\tau)$, the probability of a false rejection.

In this Thesis we will implicitly assume such assignment, according to NIST SRE evaluation protocol [Przybocki et al., 2007].

### 5.1.2. Evaluating the discriminating power

A commonly used graphical representation of the discriminating power of the detection system, especially useful when comparing multiple systems, is the ROC (Receiver -or also Relative-Operating Characteristic) plot, in which the FA rate (or $P_{fa}(\tau)$) versus the FR rate (or $P_{fr}(\tau)$) is depicted for a variable decision threshold. A variant of the ROC curve, the so-called DET (Detection Error Tradeoff) plot, is used in this Thesis [Martin et al., 1997]. In this case, the use of a non-linear scale makes the comparison of competing systems easier. A comparison between ROC and DET curves for two hypothetical competing detection systems A and B is given in Fig. 5.3.

In NIST SRE, DET plots have been used to measure the discriminating power of speaker detection technology since its early editions [Martin et al., 1997]. A summarizing measure of discriminating power is the Equal Error Rate (EER) which is the FA or FR rate at the threshold $\tau_{EER}$ where $P_{fr}(\tau_{EER}) = P_{fa}(\tau_{EER})$. The EER is easily seen as the intersection point between the DET curve and the $P_{fr}(\tau) = P_{fa}(\tau)$ line.

From the definition of discriminating power given above, it is straightforward to see that two sets of scores with the same discriminating power will have the same ROC or DET curve [Martin et al., 1997], since DET or ROC curves are just representations of the relationship among FA
5.1 Discriminating power

and FR rates for any possible threshold. It is important to remark that the actual values of
the scores are not relevant for discrimination, and two sets of scores laying at very different
numerical ranges may have the same discriminating power if the relationship between their FA
and FR rates is the same. An example is shown in Figure 5.4. There, the FA/FR curves of
two different sets of scores have been plot. The second set of scores has been obtained from
the first set of scores by an affine transformation (shifting and scaling). It is trivial to prove that,
for any threshold \( \tau \) in the first set of scores, another threshold \( \tau' \) can be found for the second
set of scores preserving the relationship between the FA and FR rates. In fact, \( \tau' \) can be easily
obtained by applying to \( \tau \) the same affine transformation which was applied to the original score
set. Thus, both sets will have the same discriminating power, and their ROC and DET curves
will be the same.

It is important to highlight that discriminating power is defined here for a given set of scores.
Thus, it is not only dependent on the automatic speaker recognition system in use, but also on
the database and protocol used for performing the trials.

5.1.2.1. Discriminating power and invertible transformations

Examples shown above demonstrate that there are some transformations of the scores which
preserve the discriminating power, e.g., an affine transformation. In fact, it can be demonstrated
that any invertible function preserves the discriminating power of a set of scores.

A function \( \phi(x) : X \to Y \) is invertible if and only if it is a bijection. If \( \phi(x) \) transforms
real numbers into real numbers, then a strictly monotonic transformation is invertible, and
therefore strictly monotonic functions preserve discriminating power as defined here (according
5. THE DISCRIMINATION LEVEL: SCORE-BASED AUTOMATIC SPEAKER RECOGNITION

![Graph showing Cumulative Distribution Function for FA and FR](image)

**Figure 5.4:** Example of FA/FR rate plots of two score sets having the same discriminating power. Score set represented in (b) has been obtained applying an affine transformation (shifting and scaling) to the scores represented in (a). Two sample thresholds (vertical lines) have been plot in (a), and their corresponding thresholds are represented in (b).

The importance of the preservation of the discriminating power by invertible transformations will be discussed in Chapter 6.

5.2. Improving discriminating power in automatic speaker recognition

In this section, we present results which illustrate several widely extended techniques for increasing the discriminating power of automatic speaker recognition systems. As it was described in Chapter 2 there are many different individual approaches in order to compute a similarity score between two speech samples. However, there are some methodologies which are common to the vast majority of state-of-the-art systems, namely:

- Score normalization.
- Session variability compensation.
- Fusion of different systems.

In the remainder of this section, we describe the state of the art in such techniques. The improvement in the discriminating power given by such techniques is illustrated by the use of several ATVS-UAM systems in the NIST SRE 2005 and 2006.
5.2 Improving discriminating power in automatic speaker recognition

5.2.1 Score normalization

Score normalization is defined as a transformation to the output scores of a speaker recognition system in order to reduce misalignments in the score ranges due to variations in the conditions of a trial. We may classify score normalization techniques into: i) training-dependent, when the variability is compensated for the training (control) speech data (also known as speaker-dependent, in reference to the known speaker who provides the training data); and ii) test-dependent, when the variability is compensated from the test (recovered) speech segment.

Many score normalization techniques have been presented in the literature, either for training- and test-dependent variability compensation [Bimbot et al., 2004]. The most popular and widely-used family of normalization techniques is the so-called impostor-centric [Bimbot et al., 2004; Fierrez-Aguilar et al., 2005b], where the normalization parameters are estimated from non-target score distributions. The reason for this popularity is two-fold: on the one hand, it is easier to obtain non-target scores rather than target scores, because the latter requires the acquisition of multiple speech utterances from a single individual, whose identity is not always known. On the other hand, the number of non-target scores which can be generated with a given database is much higher than the number of target scores, leading to more robust distributions and avoiding the effect of outliers or small sample sizes [Raudys and Jain, 1991; Theodoridis and Koutrombas, 2003]. In this Section we describe the most popular impostor-centric score normalization techniques used in automatic speaker recognition.

Z-Norm. Zero normalization technique, or Z-Norm, is derived from the work presented by Li and Porter [1988], and it has been extensively used in speaker recognition in the last ten years. A raw unnormalized score $s_{\text{raw}}$ from a trial is zero-normalized in the following way:

$$s_{\text{Znorm}} = s_{\text{raw}} - \frac{\mu_{\text{Znorm}}}{\sigma_{\text{Znorm}}}$$  \hspace{1cm} (5.1)

where $\mu_{\text{Znorm}}$ and $\sigma_{\text{Znorm}}$ are the parameters of a Gaussian distribution assigned to a set of non-target scores. This non-target score set is obtained by generating scores from the training speech used to generate $s_{\text{raw}}$ (generally represented by a speaker model) against a pool of test segments. This results in a training-dependent (or control-dependent) score distribution. The result is the alignment of the training-dependent non-target score distributions for any trial in the system. Thus, this normalization technique compensates variability in the scores due to the control speech data.

T-Norm. Test normalization, or T-Norm [Auckenthaler et al., 2000], exploits the same idea as Z-Norm, but in this case the non-target Gaussian score distribution defined by $\mu_{\text{Tnorm}}$ and $\sigma_{\text{Tnorm}}$ is assigned to sets of non-target scores generated with the test segment in each trial compared to a pool of speaker models (T-Norm cohort). The T-Norm technique is applied as follows:

$$s_{\text{Tnorm}} = s_{\text{raw}} - \frac{\mu_{\text{Tnorm}}}{\sigma_{\text{Tnorm}}}$$  \hspace{1cm} (5.2)
Thus, T-Norm performs test-dependent score normalization, and the result is the alignment of the test-dependent non-target score distributions for all trials in the system. Thus, this normalization technique compensates variability in the scores due to the recovered speech data.

**H-Norm and C-Norm.** Zero normalization can account for handset or channel variability explicitly, leading to H-Norm and C-Norm approaches respectively [Bimbot et al., 2004; Reynolds et al., 2000]. Here, the unnormalized score is mapped in the same way as Z-Norm (Equation 5.1). However, the parameters $\mu_{\text{Znorm}}$ and $\sigma_{\text{Znorm}}$ are obtained from scores generated with a pool of test segments adapted to the handset/channel of the recovered speech segment. This requires a previous automatic detection of the handset/channel of the recovered speech as described by Reynolds et al. [2000]. Therefore, this score normalization technique accounts for variabilities in the control speech sample as well as handset/channel variability in the recovered speech sample. The main drawback of this techniques rely on the need of a significant amount of normalization and handset/channel detection speech data with handset/channel labels. Such an amount of data may usually not be available.

Different combinations of the described normalization techniques have been used in the literature, leading to ZT-Norm, HT-Norm, CT-Norm, etc. Although such methods account for variability both at the control and recovered speech data, their complexity and computational demands are significantly higher. See Bimbot et al. [2004] for details.

**5.2.2. Contribution: KL-T-Norm: speaker- and test-dependent fast score normalization**

Among the score normalization techniques, T-Norm has become widely used in the automatic speaker recognition community in the last years due to its significant improvement in discrimination performance at low false acceptance rates [Navratil and Ramaswamy, 2003]. It has been shown [Reynolds, 1997; Sturim and Reynolds, 2005] that further improvement in the system can be achieved when considering both training- and test-dependent variabilities. In this section we describe a novel technique that compensates both variabilities by performing a fast selection of cohorts of models for test-normalization. The contribution of this work relies on the use of a fast approximation of the Kullback-Leibler (KL) divergence as a distance between GMM models for speaker-dependent T-Norm cohort selection. This model selection technique not only improves the system accuracy achieved using T-Norm, but also enhances the computational efficiency of the system. We have called this novel technique KL-T-Norm. This section is mainly based on Ramos-Castro et al. [2005a] and Ramos-Castro et al. [2007].

**5.2.2.1. Motivation**

The use of score normalization for simultaneously compensating test- and training-dependent variability is not new. A simple way to accomplish this objective is using different normalization
5.2 Improving discriminating power in automatic speaker recognition

Techniques simultaneously (see, e.g., [Vogt et al., 2005]). The main drawbacks of this approach are the high computational burden and the need of additional background speech sets. Moreover, the background data should be carefully selected to consider training or test conditions, and therefore these approaches can significantly increase system complexity.

Another approach to simultaneously achieve training- and test-dependent normalization is based on the incorporation of training-dependent knowledge to test-dependent techniques. One early example of such approach can be found in the context of likelihood normalization based on cohorts of speakers [Rosenberg et al., 1992]. Cohort- and test-normalization techniques are closely related, as both of them are based on the computation of the distribution of the scores that are obtained from trials using the test segment and a set of impostor models. As shown by [Reynolds, 1997] and [Finan et al., 1997], cohort normalization performance was improved when the cohorts used for normalization were different for every speaker.

Recently, a speaker-dependent T-Norm approach has been proposed in the context of test-normalization, namely adaptive T-Norm or AT-Norm [Sturim and Reynolds, 2005]. In this case, the $K$–nearest cohort models to the speaker model were used to normalize each score at each trial. A pool of utterances is compared both to the speaker and all cohort models, generating scores. A distance measure using these scores is used to select the $K$–nearest cohort models of each speaker.

The proposed KL-T-Norm technique follows AT-Norm, but we use a distance measure based on an approximation of the Kullback-Leibler divergence [Cover and Thomas, 2006] for GMM [Reynolds et al., 2000]. Our approach presents the following attractive properties:

- It is extremely fast for mean-adapted GMM from the same Universal Background Model (UBM).
- It does not require additional background data.

These properties motivate us to the use of the KL approximation described below as a distance measure to training-dependent cohort selection for test-normalization. This distance measure was presented by [Do, 2003] and has been successfully used in speaker diarization and detection [Ben et al., 2004, 2005]. Other efficient distance measures approaches for GMM may be found in [Aronowitz and Burshtein, 2007] in the context of efficient speaker identification and retrieval, and may be used as alternatives for the proposed methodology.

5.2.2.2. Test-normalization or T-Norm

The following notation is introduced for T-Norm. Let us assume that we have a sequence of feature vectors $O = \{o_1, o_2, \ldots, o_N\}$ extracted from a test utterance, and a speaker model $\lambda_t$, and that we compute a score $s(O, \lambda_t)$ by comparing the observation $O$ with the model $\lambda_t$. T-Norm uses a cohort of impostor models $\Lambda_I = \{\lambda_{I,1}, \ldots, \lambda_{I,N}\}$ to obtain the non-target scores.
5. THE DISCRIMINATION LEVEL: SCORE-BASED AUTOMATIC SPEAKER RECOGNITION

5.1 Test-Speech Features Normalization

Cohort models are used to estimate the non-target scores. Mean and Variance normalization are applied to these scores to obtain the normalized scores:

\[ s_{T\text{norm}}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{T\text{norm}}}{\sigma_{T\text{norm}}} \]  

where \( \mu_{T\text{norm}} \) and \( \sigma_{T\text{norm}} \) are respectively the mean and standard deviation of the non-target scores \( S_I \), assuming a Gaussian distribution. Figure 5.5 illustrates this process.

5.2.2.3 Fast approximation to Kullback-Leibler divergence for GMM

The Kullback-Leibler (KL) divergence between two probability density functions [Cover and Thomas, 2006] is defined by the following expression:

\[ D(f \mid \hat{f}) \equiv \int f \log \frac{f}{\hat{f}} \]  

where \( f \) and \( \hat{f} \) are arbitrary probability density functions (pdfs). KL divergence can be interpreted as an asymmetric measure of how different two probability distributions are. The solution of Equation 5.4 is classically obtained by computer-intensive algorithms such as Monte-Carlo simulation. However, these techniques usually demand high computational costs, especially when high-dimensional distributions are handled, representing a problem in real applications.

It has been recently shown in the literature [Do, 2003] that Equation 5.4 can be upper-bounded by a single expression when two Hidden Markov Models (HMM) are involved, and therefore we can particularize this expression to the GMM case. Formally, let \( f = \sum_{i=1}^{M} w_i f_i \) and \( \hat{f} = \sum_{i=1}^{M} \hat{w}_i \hat{f}_i \) be two pdfs associated with their corresponding \( d \)-dimensional GMM models \( \lambda \) and \( \hat{\lambda} \), where \( w_i \) and \( \hat{w}_i \) are real non-negative numbers (weights) and:

\[ \sum_{i=1}^{M} w_i = 1 \quad ; \quad \sum_{i=1}^{M} \hat{w}_i = 1 \]

\[ f_i = \mathcal{N} (\mu_i, \Sigma_i) \quad ; \quad \hat{f}_i = \mathcal{N} (\hat{\mu}_i, \hat{\Sigma}_i) \]  

Figure 5.5: Test-normalization technique (T-Norm).

\( S_I = \{ s(O, \lambda_{I,1}), \cdots, s(O, \lambda_{I,N}) \} \). The normalized scores are computed as follows:

\[ s_{T\text{norm}}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{T\text{norm}}}{\sigma_{T\text{norm}}} \]  

where \( \mu_{T\text{norm}} \) and \( \sigma_{T\text{norm}} \) are respectively the mean and standard deviation of the non-target scores \( S_I \), assuming a Gaussian distribution. Figure 5.5 illustrates this process.
5.2 Improving discriminating power in automatic speaker recognition

We assume a correspondence between Gaussian components of pdfs $f$ and $\hat{f}$, because in our case both distributions come from the same UBM via mean-only MAP adaptation. Without loss of generality we assume that $f_i$ and $\hat{f}_j$ are corresponding when $i = j$. Given these assumptions, we can develop the definition of the KL divergence between $f$ and $\hat{f}$ in the following way:

$$D(\frac{f}{\hat{f}}) = D\left(\sum_{i=1}^{M} w_i f_i \bigg| \sum_{i=1}^{M} \hat{w}_i \hat{f}_i\right) = \int \left(\sum_{i=1}^{M} w_i f_i\right) \log \frac{\sum_{i=1}^{M} w_i f_i}{\sum_{i=1}^{M} \hat{w}_i \hat{f}_i}$$

(5.6)

We now consider the log-sum inequality [Cover and Thomas, 2006]:

$$\left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i}$$

(5.7)

where $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ are two sets of non-negative numbers. By using Equation 5.7 the right hand side term on Equation 5.6 can be further developed as follows:

$$\int \left(\sum_{i=1}^{M} w_i f_i\right) \log \frac{\sum_{i=1}^{M} w_i f_i}{\sum_{i=1}^{M} \hat{w}_i \hat{f}_i} \leq \int \sum_{i=1}^{M} \left[w_i f_i \log \frac{w_i f_i}{\hat{w}_i \hat{f}_i}\right]$$

$$= \sum_{i=1}^{M} w_i \log \frac{w_i}{\hat{w}_i} + \sum_{i=1}^{M} w_i \int f_i \log \frac{f_i}{\hat{f}_i}$$

(5.8)

Therefore, the KL divergence between the two GMM models is upper-bounded by two terms:

- The first term is the KL divergence between the weights of both pdfs.
- The second term is the weighted sum of the individual KL divergences of the corresponding Gaussian mixtures. These individual divergences can be computed using the following formula [Do, 2003]:

$$\int f_i \log \frac{f_i}{\hat{f}_i} = \frac{1}{2} \left[ \log \frac{\det (\Sigma_i)}{\det (\hat{\Sigma}_i)} - \text{dim} (\Sigma_i) + \text{trace} \left(\Sigma_i^{-1} \hat{\Sigma}_i\right) \right]$$

$$+ \left(\mu_i - \hat{\mu}_i\right)^t \hat{\Sigma}_i^{-1} (\mu_i - \hat{\mu}_i)$$

(5.9)

\[1\] In Reynolds et al. [2000], it is shown that this adaptation scheme outperforms other MAP approaches where weights or covariances are also involved.
As our GMM are adapted from the same UBM using mean-only MAP adaptation [Reynolds et al., 2000], therefore the weight vectors and covariance matrices are the same in both models, and the first term in the right hand side of Equation 5.8 is canceled. The KL divergence approximation in this situation is reduced to the following expression:

\[
D \left( \frac{f}{\hat{f}} \right) \leq D_a \left( \frac{f}{\hat{f}} \right) = \sum_{i=1}^{M} w_i \int \frac{f_i}{\hat{f}_i} \log \frac{f_i}{\hat{f}_i} = \sum_{i=1}^{M} \frac{w_i}{2} \left[ (\mu_i - \hat{\mu}_i)^t \Sigma_i^{-1} (\mu_i - \hat{\mu}_i) \right] \quad (5.10)
\]

The approximation \( D_a \left( \frac{f}{\hat{f}} \right) \) in Equation 5.10 has several attractive properties:

**Low resource cost.** The computational cost required to compute \( D_a \left( \frac{f}{\hat{f}} \right) \) is much lower compared with other techniques for KL divergence computation such as Monte-Carlo methods.

**Symmetry.** \( D_a \left( \frac{f}{\hat{f}} \right) \) between two mean-adapted GMM from the same UBM is symmetric, i.e.,

\[
D_a \left( \frac{f}{\hat{f}} \right) = D_a \left( \frac{\hat{f}}{f} \right) \quad (5.11)
\]

**Tightness.** In [Det al., 2003], experiments show that \( D_a \left( \frac{f}{\hat{f}} \right) \) is reasonably tight to the KL divergence obtained with Monte-Carlo methods.

**Correlation.** Ben et al. [2004] show that \( D_a \left( \frac{f}{\hat{f}} \right) \) is highly correlated with the symmetric expression of the KL divergence computed via Monte-Carlo methods, namely \( D_2 \left( \frac{f}{\hat{f}} \right) = D \left( \frac{f}{\hat{f}} \right) + D \left( \frac{\hat{f}}{f} \right) \).

**Interpretation as a weighted sum of Mahalanobis distances.** The Mahalanobis distance between two multidimensional Gaussian pdfs having the same covariance matrix, namely \( g = N (\mu_g, \Sigma_g) \) and \( \hat{g} = N (\hat{\mu}_g, \Sigma_g) \) is defined as follows:

\[
D_m \left( g \mid \hat{g} \right) = \left[ (\mu_g - \hat{\mu}_g)^t \Sigma_g^{-1} (\mu_g - \hat{\mu}_g) \right] \quad (5.12)
\]

So, it can be noted that \( D_a \left( \frac{f}{\hat{f}} \right) \) is a weighted sum of Mahalanobis distances between each of the corresponding Gaussian components in each GMM.

The aforementioned properties of \( D_a \left( \frac{f}{\hat{f}} \right) \) in the model domain make it useful in areas where it may be necessary to compute distances between models, e.g., speaker diarization [Ben et al., 2004], speaker detection [Ben et al., 2005], speaker retrieval [Aronowitz and Burshtein, 2007] and the proposed score normalization.
5.2 Improving discriminating power in automatic speaker recognition

### 5.2.2.4. KL-T-Norm: speaker-dependent test-normalization

The approximation of KL divergence in Equation 5.10 is proposed to select speaker-dependent cohorts for T-Norm in automatic speaker recognition systems. We have called this novel technique KL-T-Norm. For each score \( s(O, \lambda_t) \), the application of KL-T-Norm can be described as follows:

**Computation of distances.** For each target speaker model \( \lambda_t \), we compute a set of KL divergence approximations to each model of a given cohort \( \Lambda_I = \{\lambda_{I,1}, \ldots, \lambda_{I,N}\} \), namely \( D_{t,I} = \{D_0(f_t|f_{I,1}), \ldots, D_0(f_t|f_{I,N})\} \) using Equation 5.10.

**Selection of K-nearest models.** We select the \( K \)-nearest impostor models to \( \lambda_t \) (with \( K < N \)) following the KL divergence approximation criterion, and so we will obtain a set of impostor models \( \Lambda_{KL-I} = \{\lambda_{KL-I,1}, \ldots, \lambda_{KL-I,K}\} \), being \( \Lambda_{KL-I} \subset \Lambda_I \).

**Computation of KL-T-Norm scores.** We compute the impostor scores:

\[
S_{KL-I} = \{s(O, \lambda_{KL-I,1}), \ldots, s(O, \lambda_{KL-I,K})\}
\]

**Normalization.** KL-T-Norm is finally performed as:

\[
s_{KL_T\text{norm}}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{KL_T\text{norm}}}{\sigma_{KL_T\text{norm}}}
\]

(5.13)

where \( \mu_{KL_T\text{norm}} \) and \( \sigma_{KL_T\text{norm}} \) are respectively the mean and standard deviation of \( S_{KL-I} \) assuming a Gaussian distribution.

KL-T-Norm technique is illustrated in Figure 5.6. It is important to remark that KL-T-Norm can be applied to any non-GMM speaker recognition system as well, computing \( D_{t,I} \) using GMM modeling and then use this distance set to select the KL-T-Norm cohorts. The system at hand can then be used to obtain \( s(O, \lambda_t) \) and \( S_{KL-I} \) distribution parameters in Equation 5.13.
Figure 5.7: EER performance in the development set. The number of models in the cohort, $K$, is represented in the horizontal axis for the GMM (left) and SVM-GLDS (right) systems and for 1c-1c (up) and 8c-1c (down) conditions.

Figure 5.8: minDCF performance (as defined by NIST) in the development set. The number of models in the cohort, $K$, is represented in the horizontal axis for the GMM (left) and SVM-GLDS (right) systems and for 1c-1c (up) and 8c-1c (down) conditions.
5.2.2.5. Experimental Results

In this section we present results that show that the application of KL-T-Norm to the output scores of two common speaker recognition systems improves the discrimination performance of the classical T-Norm technique. We use a GMM and a SVM-GLDS system, both of them using feature warping as channel compensation. The SVM-GLDS system uses a 2-degree polynomial kernel. Results will be presented for the NIST SRE 2005 benchmark. See Chapter 3 for details about the systems and experimental protocols.

Before the evaluation, a development set, consisting of the NIST 2004 SRE database, was selected, which is also a subset of MIXER. Trials performed using this development set follow the NIST 2004 SRE protocol \cite{vanLeeuwen2006}. Additional background data needed for development trials (UBM, normalization cohorts, etc.; namely background data) were selected from the same development set. For NIST 2005 SRE tests we used the NIST 2004 SRE database as background data. Therefore the T-Norm cohorts consist of the NIST 2004 SRE target models for each training condition. The total number of models $N$ in each cohort is shown in Table 5.1. The experiments are performed using 1 conversation side for testing and both 1 and 8 conversation sides for training (1c-1c and 8c-1c conditions respectively).

Figures 5.7 and 5.8 summarize the experiments performed in the development set. We vary the number of models $K$ used for KL-T-Norm and plot the EER in (%), and we can compare the use of T-Norm ($K = N$) and several operating points of KL-T-Norm ($K = 25, ..., 150$). We observe that KL-T-Norm improves the system performance in terms of Equal Error Rate (EER) especially for $K = 50$ and $K = 75$. We also observe that the optimum EER value is obtained for $K = 50$. Figure 5.8 shows the optimum Detection Cost Function (minDCF) as defined by NIST evaluations \cite{Przybocki2007} for the same experiments as in Figure 5.7. The minDCF is also a measure of discriminating power, and it will be introduced in Chapter 6. We observe that a general trend of minDCF improvement exists for KL-T-Norm in all cases.

In order to perform KL-T-Norm in NIST 2005 SRE, we set $K = 75$ for both systems (GMM and SVM-GLDS) and conditions (1c-1c and 8c-1c). Figure 5.9 shows the performance of the GMM and SVM-GLDS systems in both 1c-1c and 8c-1c conditions when no normalization, T-Norm and KL-T-Norm are used. For all these cases, T-Norm has been applied using the total number of models in the cohorts (Table 5.1). We note an improvement in system performance for KL-T-Norm with respect to T-Norm in the 8c-1c condition, whereas this is not appreciated for the 1c-1c condition.

<table>
<thead>
<tr>
<th>$N$ in cohort</th>
<th>1c-1c</th>
<th>8c-1c</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>246</td>
<td>170</td>
</tr>
<tr>
<td>female</td>
<td>370</td>
<td>205</td>
</tr>
</tbody>
</table>

Table 5.1: Total number of models $N$ in each T-Norm cohort

The efficiency gain of KL-T-Norm can be illustrated by estimating the processing time of
T-Norm and KL-T-Norm:

\[ t_{T\text{Norm}} = N \cdot t_{\text{score}} \]
\[ t_{KL-T\text{Norm}} = N \cdot t_{KLa} + K \cdot t_{\text{score}} \]

where \( t_{\text{score}} \) is the time needed to compute a score using the fast scoring technique presented by Reynolds et al. [2000] and \( t_{KLa} \) is the time required to perform \( D_a \) using Equation 5.10.

Defining \( R = t_{\text{score}} / t_{KLa} \) we obtain:

\[ t_{KL-T\text{Norm}} = t_{T\text{norm}} \left( \frac{1}{R} + \frac{K}{N} \right) \]  

(5.15)

Note that if \((1/R + K/N) < 1\) then we obtain a computational gain. We have estimated from our experiments that these values are, on average, \( R \approx 10 \) and \( K/N \approx 0.3 \). Therefore the computational cost of the score normalization technique is decreased to \( t_{KL-T\text{Norm}} \approx 0.4 \cdot t_{T\text{norm}} \).

In order to compare KL-T-Norm to T-Norm at similar operating points in terms of computational burden, we compare here both of them with the same number of models in the cohorts. As before, we use \( K = 75 \) models in KL-T-Norm for all experiments. In an analogous way, we apply T-Norm using \( K = 75 \) models randomly selected from each whole cohort. In order to perform statistically significant experiments, we evaluate the system using 10 different random selections for T-Norm cohorts, averaging the results. Table 5.2 for the GMM system shows a significant EER improvement in all conditions when KL-T-Norm is used. On the other hand, a slight improvement in minDCF values is appreciated in all cases. Table 5.3 shows the same results for the SVM-GLDS system. In this case the performance gain is even higher.

---

Footnote:

1We have empirically seen that the rest of processing times (sorting the distances, memory access, etc.) are negligible compared to \( t_{KLa} \) and \( t_{\text{score}} \).
5.2 Improving discriminating power in automatic speaker recognition

### Table 5.2: T-norm and KL-T-norm in GMM system using $K = 75$ models for both techniques (T-Norm values are averaged from 10 random selection trials)

<table>
<thead>
<tr>
<th></th>
<th>1c-1c</th>
<th>8c-1c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>male</td>
<td>female</td>
</tr>
<tr>
<td>EER T-Norm (Av.)</td>
<td>11.14</td>
<td>14.62</td>
</tr>
<tr>
<td>EER Av. Improvement</td>
<td>3.4%</td>
<td>5.0%</td>
</tr>
<tr>
<td>minDCF T-Norm (Av.)</td>
<td>0.041</td>
<td>0.048</td>
</tr>
<tr>
<td>minDCF KL-T-Norm</td>
<td>0.039</td>
<td>0.047</td>
</tr>
</tbody>
</table>

### Table 5.3: T-Norm and KL-T-Norm in SVM-GLDS system using $K = 75$ models for both techniques (T-Norm values are averaged from 10 random selection trials)

<table>
<thead>
<tr>
<th></th>
<th>1c-1c</th>
<th>8c-1c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>male</td>
<td>female</td>
</tr>
<tr>
<td>EER T-Norm (Av.)</td>
<td>19.22</td>
<td>19.27</td>
</tr>
<tr>
<td>EER KL-T-Norm</td>
<td>17.19</td>
<td>17.87</td>
</tr>
<tr>
<td>EER Av. Improvement</td>
<td>11.6%</td>
<td>7.3%</td>
</tr>
<tr>
<td>minDCF T-Norm (Av.)</td>
<td>0.073</td>
<td>0.075</td>
</tr>
<tr>
<td>minDCF KL-T-Norm</td>
<td>0.063</td>
<td>0.073</td>
</tr>
</tbody>
</table>

5.2.3. Session variability compensation

One of the main problems in automatic speaker recognition comes from the variability in the conditions of different speech utterances. The speech signal is affected at all levels by the transmission channel, the reverberation in a room, the ambient noise, the speaking style, the gender of the speaker, the effect of aging or merely the temporal separation between recording sessions [Castaldo et al., 2007; Kenny et al., 2007; Vogt and Sridharan, 2007]. Different conditions in recovered and control speech utterances greatly affect the discrimination performance of the scores computed by the speaker recognition system. It then seems straightforward that a fundamental success of speaker recognition technologies has come from session variability compensation, according to Reynolds [2003b].

Most successful compensation techniques have been proposed at the spectral level.

Numerous compensation techniques for channel effects have been proposed and applied in three domains:

- **Feature domain**, where compensation aims at removing degrading variability from the feature vectors prior to speaker model building or score computation.

- **Model domain**, where the models of the features are compensated to minimize the effects of
Score domain, also known as score normalization, and previously described in this section.

Compensation at different domains is not exclusive. Indeed, compensation at each domain deals with a different aspect of session variability, and complementary benefits can be expected from the combination of techniques.

There have been many approaches for session variability compensation, working at the feature, model (e.g., GMM or SVM) or score domain. First attempts worked at the feature level with cepstral parameters. Cepstral representation has the advantage that invariant channels add a constant offset that can be easily subtracted. This technique is known as Cepstral Mean Subtraction (CMS) or Normalization (CMN). Moreover, the analysis of the variation of cepstral features for speech and non-speech features led to the development of RASTA filtering of cepstral instantaneous vectors. These techniques have been widely used in speaker recognition in the last decades. One of the main drawbacks of these techniques is the simplicity of the approach, which does not consider many other variability sources. Moreover, these techniques only address variability affecting each dimension of the feature vectors separately, and variability in the correlation between cepstral coefficients is not considered.

Feature warping was proposed for channel compensation at the feature vector level. The histogram of each cepstral coefficient over time was matched to a given probability distribution. This is motivated because it was observed that environmental effects such as noise or transmission channel affected the distribution of the cepstral parameters rather than its temporal structure. In the target distribution was a standardized Gaussian (zero mean, variance unity), though the authors motivated the use of more sophisticated distributions. Again, feature warping does not consider the correlation between coefficients. However, its application is simple and it works reasonably well without the need of additional speech data for warping training.

Speaker Model Synthesis (SMS) was proposed to compensate variability in the models in GMM-based systems. The approach aims at transforming the GMM trained from control speech in order to adapt it to the conditions of the recovered speech. This technique considers correlation between features, as the transformation is applied in the model space, and not independently to each coefficient. However, it has three main drawbacks. First, the technique can only be applied to GMM models. Second, it needs a considerable amount of training data with accurate training labels in order to estimate the transformations to apply to the GMM depending on the conditions of the recovered speech. Third, it assumes that session variability is produced by a discrete set of degrading factors, while it may be many other latent sources of variability in the speech signal.

Part of the problems of SMS were solved by feature mapping. While SMS applied a channel-dependent transformation to the GMM models, feature mapping proposed applying the inverse transformation to the feature vectors. That resulted in a space with
5.2 Improving discriminating power in automatic speaker recognition

Compensated conditions where all the features were mapped to, which allowed the use of such features with any other modeling technique apart from GMM, such as SVM-GLDS, SVM-GMM-SV, etc. However, as SMS, feature mapping still required a significant amount of training data with channel labels in order to work properly, and it also still assumes session variability as produced by a set of discrete factors.

Recently, a significant advance in the state of the art has been introduced by the use of factor analysis techniques for session variability compensation [Castaldo et al., 2007; Kenny and Dumouchel, 2004; Vogt and Sridharan, 2007]. This technique models the maximum variability directions due to speaker and inter-session effects, and aims at compensating the latter while preserving the former. While the original approach was applied to GMM models [Kenny et al., 2007], the superior performance of these techniques has motivated the upcoming of many variants, working at the feature [Vair et al., 2006] or at the SVM models domain (Nuisance Attribute Projection [Solomonoff et al., 2005]). Moreover, these techniques do not need any labels indicating the conditions of the speech data and any assumed set of degrading factors, since latent sources of variability of the feature vectors are naturally modeled by factor analysis. Top-performing short-time spectral level systems (GMM, GMM-SVM-SV, MLLR-SVM) together with session variability compensation based on factor analysis lead the state-of-the-art of individual automatic speaker recognition systems.

5.2.3.1. Session variability compensation experiments

This section illustrates the effect of inter-session variability compensation in the discriminating power of automatic speaker recognition systems. Experiments will focus on the comparison of feature warping and feature mapping to the recently proposed Nuisance Attribute Projection (NAP) technique [Solomonoff et al., 2005]. Experiments shown here follow the NIST SRE 2006 database and protocol for the 1conv4w-1conv4w test condition. Details about baseline systems and experimental protocols can be found in Chapter 3.

First, the effects of NAP in a GMM-SVM-Supervectors system with feature mapping is presented. Results in Figure 5.10(a) show the significant improvement in the discriminating power of the system if NAP is performed in order to compensate session variability. Moreover, as NAP compensates variability at the GMM model level, it is possible to use NAP-compensated models for a GMM system. Results of such compensation scheme are shown in Figure 5.10(b) for a GMM system with feature warping. It is also shown a remarkable improvement in the discrimination performance.

5.2.4. Fusion

Given the multi-level and multi-system nature of automatic speaker recognition (Chapter 2) the combination of outputs coming from different systems for a given trial arises naturally. In fact, the state of the art in automatic speaker recognition is dominated by the combination of multiple systems exploiting different levels of information of the speaker identity, mainly at the
spectral level [Brümmer et al., 2007]. This combination paradigms are known as fusion techniques, and have been widely used in biometrics in the last years, yielding abundant theoretical frameworks and practical methods in the biometrics literature [Jain and Ross, 2004].

In this Thesis, we will concentrate on several fusion strategies used in automatic speaker recognition. In this area, top-leading fusion schemes are limited to the combination or classification of the scores coming from different systems [Brümmer et al., 2007; Fierrez, 2006; Garcia-Romero et al., 2003; Pigeon et al., 2000]. Therefore, we classify the following fusion approaches into multiple systems combination and multiple systems classification fusion.

In multiple systems combination, fusion is reduced to simple operators such as the product, sum or average of the output scores of each system. Kittler et al. [1998] followed this approach in a probabilistic framework and provided an example of multimodal biometric authentication fusing speech, frontal and profile images modalities. In automatic speaker recognition this approach has been always simplified to score combination using such simple operators, where the fused score is obtained usually by a sum (sum fusion), or a weighted average of the input scores [Campbell et al., 2004c]. As arithmetic or logical rules are fixed, the resulting scores will be biased to one of the combining systems if they are not properly normalized to the same range prior to fusion [Campbell et al., 2004c; Jain et al., 2004; Jain and Ross, 2004]. Thus, score normalization is essential in order to perform multiple systems combination fusion.

On the other hand, multiple systems classification fusion is based on considering the fusion stage as a second-level pattern recognition problem [Duda et al., 2001]. In this case the output scores from the different systems are considered components of a new feature vector which is the input to a second-level classifier. Thus, within this fusion strategy, any of the classifiers available from the literature can be used. In automatic speaker recognition, support vector machines
5.2 Improving discriminating power in automatic speaker recognition

(SVM) Garcia-Romero et al. [2003] and logistic regression Brümmer et al. [2007; Pigeon et al. 2000] have been mainly used for classification fusion. The former presents the typical robustness of SVM classifiers. The latter is computationally efficient and presents desirable properties which will be discussed later in Chapter 5.

Adapted fusion schemes have been also proposed for speaker recognition, both in multilevel speaker recognition Fierrez-Aguilar et al. [2005a] or in combination to other biometric traits (multimodal fusion) Jain and Ross [2004]. There, the classification fusion rules, based on SVM or Bayesian classifiers, were adapted to the speaker specificities or the quality of the input signal, resulting in a significant performance improvement in some scenarios.

5.2.4.1. Fusion experiments

In this section, we show the effect of the fusion of different automatic speaker recognition systems at the spectral and higher levels. Experiments are conducted using NIST SRE 2005 and 2006 datasets. Moreover, two test conditions are presented in all cases, namely 8conv4w-1conv4w and 1conv4w-1conv4w. In the former, the amount and diversity of training data (8 different conversation sides of approximately 2.5 minutes each) allows the efficient use of high-level speaker recognition systems (phonetic and prosodic). In the latter, spectral systems are fused, as the performance of higher level systems is poor. Details about baseline systems and experimental protocol can be found in Chapter 3.

First, we show results of fusion of the ATVS-UAM GMM and SVM-GLDS spectral systems in the 1conv4w-1conv4w condition. For the SVM-GLDS system, 3–degree polynomial expansion has been used. Also, SVM-GLDS system works in two configurations, namely forward and reverse, where the only difference is that the reverse configuration uses the training (control) speech data as testing (recovered) speech data and vice-versa. Both systems use the same feature extraction scheme, with feature mapping for channel compensation.

A simple sum fusion scheme has been used for NIST SRE 2005, whereas this simple method is compared to logistic regression fusion in NIST SRE 2006 experiments. Scores from the NIST SRE 2005 have been used in order to train the logistic regression fusion for NIST SRE 2006. Figure 5.11 shows the results for both benchmarks. Simple sum fusion significantly improves the performance of the combined system in NIST SRE 2005, especially at low FA rates, although a slight degradation is appreciated in some areas around the EER. However, in NIST SRE 2006 sum fusion has roughly the same discriminating power as the best-performing GMM system. On the other hand, logistic regression fusion increases the discriminating power of the best system, but only slightly. This fact is mainly due to two factors. First, the three presented systems use the same feature extraction scheme, and therefore it is more difficult to find complementary information among them. Second, NIST SRE 2006 database presents a higher variability than NIST SRE 2005 for the 1conv4w-1conv4w task, mainly due to the inclusion of additional multi-language data. Thus, the score normalization process is more difficult for the SRE 2006 corpus, and the performance of the sum fusion degrades.

We now show results of the fusion of spectral and higher level systems for the 8conv4w-
Figure 5.11: Effects of the fusion of ATVS GMM and SVM-GLDS systems for the 1conv4w-1conv4w condition of NIST SRE 2005 (a) and NIST SRE 2006 (b).

Figure 5.12: Effects of the fusion of ATVS-UAM spectral and higher-level systems for the 8conv4w-1conv4w condition of NIST SRE 2005 (a) and NIST SRE 2006 (b).
5.3 Chapter summary and conclusions

In this case, the GMM and SVM-GLDS systems (in its forward configuration) are fused with a prosodic system and three phonetic systems. The prosodic system is based on pitch and energy slope tokenization, whereas the phonetic system is based on PRLM n-gram modelling using three phonetic recognizers trained with speech in English, Spanish and Basque respectively. Details about high-level systems can be found in Chapter 2. Figure 5.11 shows the fusion results for both systems. It is shown that in this case sum fusion is not exploiting the complementary information given by the different systems. The main reason for this is due to correlation: phonetic systems are strongly correlated, and their performance is poorer than for spectral systems. Then, as sum fusion is associative and it gives the same importance to all systems, many correlated systems with poorer performance will degrade the overall fusion performance. Moreover, because of the different nature of the spectral and higher level systems, it is difficult to find a proper alignment for sum fusion to work, even if T-Norm is used for score normalization. On the other hand, logistic regression fusion in NIST SRE 2006 significantly increases the discriminating power, demonstrating the complementary nature of spectral and higher level systems. As logistic regression takes correlation of systems into account [Pigeon et al., 2000], the degrading effect of poorer correlated systems is not observed.

5.3. Chapter summary and conclusions

In this chapter we have described the discrimination abstraction level in the proposed hierarchical methodology for forensic automatic speaker recognition. First, the aim at this level has been defined as the generation of a similarity score between the control and recovered speech material, which will allow the discrimination between same-source and different-source trials. The inputs needed for this purpose and the outputs generated at this level are presented as well. A performance measure has been described for this abstraction level, namely discriminating power (or discrimination performance), which represents the ability of the automatic speaker recognition system in order to separate same-source trials from different-source trials.

The chapter continues with experimental results illustrating several techniques which have been extensively used for increasing the discriminating power of state-of-the-art systems, namely score normalization, session variability compensation and fusion of individual systems. The novel KL-T-Norm technique has been introduced as an efficient method for score normalization considering the variability of both the training and the test speech data. Experimental results using several ATVS-UAM systems have been reported for different NIST SRE protocols and test conditions.

The original contribution in this chapter is the novel KL-T-Norm score normalization technique.
Chapter 6

The presentation level: $LR$ computation

This chapter introduces the presentation abstraction level in the hierarchical methodology for forensic automatic speaker recognition proposed in this Thesis. The aim at this level is presenting the discriminating information carried by the score in a proper way, in order to give the fact finder meaningful information about the weight of the forensic evidence.

In the proposed methodology, $LR$ values will be used to present the weight of the evidence in court. This follows the DNA methodology, as it has been justified in Chapter 4. However, it is unrealistic to pretend that the $LR$ computation process will be performed with a perfect knowledge about the problem. In fact, variability of the speech signal and speech data scarcity are typical in forensic applications, and therefore $LR$ computation will be a complicated task in general. Thus, there is a need of measuring not only the information given by the discriminating power of the score, but also the ability of presenting such results without loss of information. Such a measure of goodness will state the accuracy of the forensic automatic speaker recognition system on a given experimental test. As remarked in Chapter 4, the definition of accuracy is one of the objectives of the proposed methodology.

The presentation level is integrated into the proposed hierarchical methodology for speaker recognition, and it has been defined in Chapter 4. In Figure 6.1, the inputs and outputs of this level are detailed. On the one hand, this level receives the following inputs:

- **Similarity score (from the discrimination level).** The score computed by the automatic speaker recognition system carries the information about whether the control and recovered speech material come from the same source or not. This score will be transformed at this level in order to properly present such information to the fact finder, following the $LR$ methodology.

- **Discrimination performance (from discrimination level)** of the scores generated by the forensic automatic speaker recognition systems is essential at the presentation level.
6. THE PRESENTATION LEVEL: LR COMPUTATION

The score-to-\( LR \) transformation should not degrade the discriminating power given by the input score.

- **Background information about the case.** The computation of the \( LR \) value requires the selection of a proper population of individuals in order to model the variabilities involved, as well as the use of carefully selected databases depending on the conditions of the recordings to be analyzed. Therefore, it is important to have knowledge about background information such as the recording conditions, the suspect native language or social origin, etc.

- **Speech databases.** Speech data is needed to model populations, within- and between-speaker variations, etc.

On the other hand, the presentation level yields the following outputs, which will be inputs for the next level (the forensic level, described in Chapter 7):

- **LR value.** This value represents the weight of the evidence in the forensic case.

- **Testing results.** The results from a proper evaluation of the \( LR \) value should be submitted to the next level in order to transparently state the accuracy of the technique in

---

**Figure 6.1:** The presentation level in the proposed hierarchical methodology for forensic automatic speaker recognition.
6.1 Evaluation of the evidence using LR values

This chapter is organized as follows. First, the LR methodology for evidence evaluation is addressed. We then discuss the meaning of a LR value and its role in the Bayesian inferential process, in order to highlight the importance of the LR computation process. Also, Bayesian decision theory is described, as a theoretical framework for making optimal decisions. Then, previously proposed measures of the goodness of the LR are introduced. We then discuss the important concept of calibration with several illustrating examples, and then we define what is understood as the accuracy of an LR–based speaker recognition system. A novel performance representation for measuring the accuracy of LR values is then presented, based on the information-theoretical concept of empirical cross-entropy (ECE). This representation, namely the ECE plot, generalizes existing methods within an information-theoretical framework. The Dissertation then focuses on the obtention of accurate LR values from the scores delivered by the discrimination level. A classification of LR computation methods is presented, and different approaches for obtaining accurate LR values present in the literature will be described and compared. As a contribution, a novel suspect-adapted LR computation technique is proposed, with an experimental section which demonstrates its adequacy. Conclusions are finally drawn.

Original contributions in this chapter include the information-theoretical framework for the definition, interpretation and assessment of the accuracy of LR values; and the novel suspect-adapted LR computation technique.

6.1. Evaluation of the evidence using LR values

The LR framework for interpretation of the evidence represents a mathematical and logical tool in order to aid in the inference process derived from the analysis of the evidence. In this methodology, the objective of the forensic scientist is computing the likelihood ratio (LR) as a degree of support of one hypothesis versus its opposite [Aitken and Taroni, 2004; Champod and Meuwly, 2000].

The LR framework is stated as follows. Consider the forensic speech evidence E as the comparison of a recovered speech sample of unknown origin and a control sample whose origin is known. In a forensic case, the unobserved variable of interest is the true hypothesis \( \theta = \{ \theta_p, \theta_d \} \), where:

- \( \theta_p \) means that the suspect is the source of the speech material of disputed origin.
- \( \theta_d \) means that another individual in the relevant population is the source of the speech material of disputed origin.

Bayes’ theorem [Aitken and Taroni, 2004; Papoulis, 2001] relates probabilities before and after evidence analysis:

\[
P(\theta_p \mid E, I) = \frac{p(E \mid \theta_p, I) \cdot P(\theta_p \mid I)}{p(E \mid I)}
\]  
(6.1)
where $I$ is the background information available in the case not related to the evidence $E$. This $I$ may include not only circumstantial information in the case (such as witness testimony), but also the analysis of other forensic evidence apart from $E$ (such as, glass fragments, paint flakes, etc.). Equation (6.1) then allows the following inference:

$$\frac{P(\theta_p \mid E, I)}{P(\theta_d \mid E, I)} = LR \frac{P(\theta_p \mid I)}{P(\theta_d \mid I)}$$

(6.2)

$$LR = \frac{p(e \mid \theta_p, I)}{p(e \mid \theta_d, I)}_{e=E}$$

(6.3)

Equation (6.2) is the so-called odds form of Bayes’ theorem. The hypotheses should be defined in the court from $I$, the prosecutor and defense propositions and often because of the adversarial nature of the criminal system. In this framework, we can distinguish two values:

1. The prior probabilities $P(\theta_p \mid I) = 1 - P(\theta_d \mid I)$, which are province of the fact finder and should be stated assuming only the background information ($I$) in the case [Evett, 1998].

2. The $LR$ (Equation (6.3)), computed by the forensic scientist [Aitken and Tarotni, 2004].

The $LR$ value (Equation (6.3)) is the quotient of two probability densities. On the one hand, the probability density function (pdf) $p(e \mid \theta_p, I)$ in the numerator in Equation (6.3) is known as the within-source distribution, and models the variability of the speaker between sessions. Its evaluation in $e = E$ gives a measure of the similarity between the recovered and control materials. On the other hand, the pdf $p(e \mid \theta_d, I)$ in the denominator is known as the between-source distribution, and its evaluation in $e = E$ can be seen as a measure of the typicality or rarity of the recovered material in the relevant population. Both values, similarity and typicality, should be computed in a transparent way by the forensic scientist. It is also the duty of the forensic scientist, following the background information of the case ($I$), to select the population of individuals which will be proper for the case at hand.

This $LR$–based framework presents many advantages in a forensic context:

- It allows forensic scientists to evaluate and report a meaningful value for the weight of the evidence to the court [Champod and Meuwly, 2000].

- The role of the scientist is clearly defined, leaving to the court the task of using prior judgments or costs in the decision process.

- Probabilities can be interpreted as degrees of belief [Taroni et al., 2001], allowing the incorporation of subjective opinions as probabilities in the inference process in a clear and scientific way.

1Unless explicitly stated, we will use a capital $E$ for referring to the given value of the evidence, according to the literature on $LR$–based analysis of the evidence [Aitken and Tarotni, 2004; Champod and Meuwly, 2000; Drygajlo, 2002]. Thus, the small $e$ will be used as the argument in likelihoods.

2The background information about the case $I$ will be eliminated from the notation for the sake of simplicity from here thereafter. It will be assumed that all the probabilities defined are conditioned to $I$. 

88
6.1 Evaluation of the evidence using LR values

6.1.1. The meaning of the LR

In order to interpret the meaning of the LR value, it is useful to express probabilities about \( \theta \) in terms of odds, defined as the quotient between the probability of an event against the probability of its complementary event. For instance, \( \theta_p \) and \( \theta_d \) are complementary events, because they are mutually exclusive and exhaustive. Thus, the prior odds for \( \theta_p \) are defined as:

\[
O(\theta_p) = \frac{P(\theta_p)}{P(\theta_d)} = \frac{P(\theta_p)}{1 - P(\theta_p)}
\]

and the posterior odds for \( \theta_p \) are defined as:

\[
O(\theta_p|E) = \frac{P(\theta_p|E)}{P(\theta_d|E)} = \frac{P(\theta_p|E)}{1 - P(\theta_p|E)}
\]

The odds in favor of any of the hypotheses \( \theta_p \) and \( \theta_d \) can be referred to as the opinion about whether the unknown hypothesis value \( \theta \) is \( \theta_p \) or \( \theta_d \), conditioned to the available information. They also have an intuitive interpretation. For instance, on the one hand \( O(\theta_p) = 2 \) expresses that the opinion is 2 versus 1 in favor of \( \theta = \theta_p \). On the other hand, \( O(\theta_p) = \frac{1}{3} \) means that the opinion is 3 versus 1 in favor of \( \theta = \theta_d \). An intuitive metaphor can be established if we see odds as stones in a balance, as it is illustrated in Figure 6.2. It is easily seen from Equation 6.4 that \( O(\theta_p) = \frac{1}{O(\theta_d)} \) and that \( O(\theta_p|E) = \frac{1}{O(\theta_d|E)} \).

Prior to the evaluation of the evidence, the available information is given solely by \( I \), the background information of the case. After evidence evaluation, the value of the evidence \( E \) is also included into the available information, and the odds (opinion about the true value of the hypothesis variable \( \theta \)) are subsequently modified. Posterior odds represent the opinion of the fact finder in the light of all the available knowledge about the problem, namely \( E \) and \( I \). The relationship between the prior and posterior odds is derived from Equation 6.2 and in the form of odds is given by:

\footnote{The author believes that the balance is an especially illustrating metaphor in a forensic context.}

\footnote{Remember that \( I \) was eliminated from the notation for simplicity, but all probabilities involved in the case are assumed to be conditioned to \( I \).}
6. THE PRESENTATION LEVEL: LR COMPUTATION

Prior odds

\[ \text{Before evidence analysis} \]

\[ \theta_p, \theta_d \]

LR

\[ \text{Weight of the evidence} \]

\[ LR = 10 \]

\[ LR = 10 \]

\[ LR = 100 \]

Posterior odds

\[ \text{After evidence analysis} \]

\[ \theta_p, \theta_d \]

\[ 5, 1 \]

\[ 20 \]

\[ 5, 1 \]

\[ 20 \]

\[ 5, 1 \]

Case a)

Case b)

Case c)

\[ O(\theta_p|E) = LR \cdot O(\theta_p) \quad (6.6) \]

According to Equation 6.6, the value which represents a change in opinion due to new information is the \( LR \), i.e., the weight of the evidence \[ \text{Aitken and Taroni, 2004} \]. The role of the \( LR \) value in the inferential process is illustrated in Figure 6.3. There, the prior odds are modified by the \( LR \) value to obtain the posterior odds. After observing Figure 6.3, several remarks should be considered:

- In two different cases, the same \( LR \) value for both cases may lead to different posterior odds, because the prior odds may be different. That represents that the same weight of the evidence (the same \( LR \) value) will lead to a different opinion after the analysis of the evidence if the opinion before the analysis of the evidence was sufficiently different. This is illustrated in cases (a) and (b) in Figure 6.3.

- In two different cases, different \( LR \) values may lead to the same posterior odds even if the prior odds are different. That represents that the opinion after the analysis of the evidence may be the same for two cases, even if the opinions before the analysis of the evidence were very different. This is illustrated in cases (a) and (c) in Figure 6.3.

Finally, in order to obtain posterior probabilities, the following relationship can be derived from Equations 6.5 and 6.6:

\[ P(\theta_p|E) = \frac{LR \cdot O(\theta_p)}{1 + LR \cdot O(\theta_p)} \quad (6.7) \]
6.1 Evaluation of the evidence using LR values

6.1.1.1. Interpretation of the LR

From the previously described analysis and examples, it seems clear that the LR value has an interpretation as a support to a previously stated opinion, due to the analysis of the evidence E. In other words:

- If the $LR > 1$ the evidence will support that $\theta = \theta_p$, i.e., it will support that the control and the recovered speech material come from the same source.

- If the $LR < 1$ the evidence will support that $\theta = \theta_d$, i.e., it will support that the control and the recovered speech material come from different sources from the relevant population.

Moreover, the value of the LR represents the degree of support of the evidence to the value of $\theta$. For instance, $LR = 3$ meaning “the evidence supports the odds in favor of $\theta = \theta_p$ with a degree of 3”. Therefore, a single LR value has a meaning by itself.

It is important to note that the LR supports an opinion about $\theta$, but the LR is not an opinion about $\theta$. Opinions about $\theta$ are represented as probabilities or, in our binary case, odds in favor of a given outcome of $\theta$. Therefore, it is not possible to make a decision about the value of $\theta$ based solely on the value of the LR, because decisions will be taken from posterior opinions as it will be shown later in this chapter.

6.1.1.2. LR values and similarity scores

For the subsequent discussion, it will be useful to remark the main difference in the concepts of similarity score and LR value.

As described in previous chapters, a similarity score between a pair of control and recovered speech material is a real value which is bigger for higher support to $\theta = \theta_p$, and the smaller the score, the greater the support to $\theta = \theta_d$. However, it is not possible in general to state the degree of support of a single similarity score. For instance, a score value of 3 for a given control-recovered comparison does not mean anything by itself if no other information is given (e.g., a decision threshold, the distribution of other target and non-target scores, etc.).

Therefore, the similarity score cannot be used as a LR value in general. Hence, if a score is intended to be used in the same sense as a LR value in Equation 6.6, the posterior odds inferred may be highly misleading.

The definition of a score given here implies that the LR can be used as a similarity score, because the support to $\theta_p$ raises as the LR value is bigger, and vice-versa. Therefore, a LR value may be used as a discriminating value as described in Chapter 5, and it is possible to compute the discriminating power of a set of LR values (e.g., through a DET plot).

In conclusion, the main difference between a score and a LR value relies upon its functionality. Both the score or the LR value may act as discriminating values, separating same-source and different-source comparisons. However, the LR value will be also intended to be used to obtain posterior odds by the use of Equation 6.6.
Through all this Dissertation, scores and \( LR \) values will be compared in order to highlight several effects derived from their use in different frameworks. It is important to remark that, when a comparison between \( LR \) values and scores is be performed through the dissertation, the domain of comparison will be the same. Therefore, if the scores vary in the \((0, \infty)\) interval, they will be compared to \( LR \) values, which also vary in the interval \((0, \infty)\). On the other hand, if the range of variation of the scores is \((-\infty, \infty)\), they will be compared to \( \log(LR) \) values in a logarithmic domain. The latter is the common situation in automatic speaker recognition, since most systems output scores varying between \(-\infty\) and \(\infty\).

6.2. Bayesian decision theory

The ultimate aim of forensic identification belongs to the fact finder, who should take the final decision about the innocent or guilt of a given suspect. Bayesian decision theory gives the mathematical tool for making optimal decisions in the light of evidence. In this sense, Bayesian decision framework is a step further from Equation 6.2: if the fact finder has the posterior probabilities, he will be able to make a decision. Bayesian decision theory gives the optimal solution when there is plenty of information about the probabilities involved in the problem.

In a forensic framework, the control-recovered material pair should be classified by the fact finder as coming from the same source \( (\theta_p) \) or as coming from different sources belonging to the relevant population \((\theta_d)\). The Bayes decision rule is stated as follows:

For a given value of \( e = E \):

\[
\begin{align*}
\text{decide } \theta_p & : P(\theta_p | E) > P(\theta_d | E) \\
\text{decide } \theta_d & : P(\theta_p | E) < P(\theta_d | E)
\end{align*}
\]

\( (6.8) \)

Following Bayes’ theorem (Equation 6.1), it is easily shown that the Bayes decision rule can be also stated as:

For a given value of \( e = E \):

\[
\begin{align*}
\text{decide } \theta_p & : p(E | \theta_p) \cdot P(\theta_p) > p(E | \theta_d) \cdot P(\theta_d) \\
\text{decide } \theta_d & : p(E | \theta_p) \cdot P(\theta_p) < p(E | \theta_d) \cdot P(\theta_d)
\end{align*}
\]

\( (6.9) \)

and therefore:

\[
\begin{align*}
\text{decide } \theta_p & : LR > \frac{P(\theta_d)}{P(\theta_p)} \\
\text{decide } \theta_d & : LR < \frac{P(\theta_d)}{P(\theta_p)}
\end{align*}
\]

\( (6.10) \)

i.e., the decision is taken comparing the \( LR \) value to a threshold \( \tau \) which depends on the value of the priors. Figure 6.4 illustrates the definition of the decision boundaries from the \( LR \) value and the priors, according to Equation 6.10.

It can be demonstrated that the Bayes decision rule minimizes the probability of error [Duda et al., 2001; Theodoridis and Koutroumbas, 2003]. There are two kind of errors in this binary context:
6.2 Bayesian decision theory

Figure 6.4: Definition of the decision regions ($R_p$ and $R_d$) in Bayesian decision theory. (a) shows the value of the likelihoods of $e$ conditioned to $\theta_p$ and $\theta_d$ respectively. The boundary $B$ of the decision regions is placed according to the value of the likelihoods and Bayes decision rule (Equation 6.8), being $R_p$ the region of the $e$ space where $\theta_p$ is decided, and vice-versa. (b) shows the relationship between the decision boundary in the $e$ space and the logarithm of the LR. $\theta_p$ will be decided if $LR > \frac{P(\theta_p)}{P(\theta_d)}$ and vice-versa. In this example, ($P(\theta_p) = P(\theta_d)$) and therefore the logarithm of the decision threshold is zero.

Figure 6.5: Definition of the decision regions ($R_p$ and $R_d$) in Bayesian decision theory when Bayes’ threshold $\tau_B$ (priors and costs) change. In (a) we show the likelihoods and decision regions in the $e$ space. If the decision boundaries are chosen according to Bayes decision rule (Equation 6.13) then an increase in $P(\theta_p)$ or in $C_{fa}$ (or in both) will move the decision boundary in order to restrict the $R_d$ region (boundary $B'$). Conversely, if $P(\theta_d)$ or $C_{fa}$ (or both) increase then the boundary will move in order to restrict $R_p$. In (b) it is shown that a restriction in $R_d$ (boundary $B'$) corresponds to a lower Bayes threshold $\tau'$ for the LR, and vice-versa.
The fact finder decides $\theta_d$ and the true value of the hypotheses is $\theta_p$. This kind of error is known as a *false negative*, a *type I error* or a *false rejection* (FR) [Aitken and Taroni 2004; Champod and Meuwly 2000]. If this error occurs, the fact finder will state that the control and recovered materials come from different sources, but they actually come from the same source. Roughly speaking, the fact finder will falsely decide towards the innocence of a guilty person.

The fact finder decides $\theta_p$ and the true value of the hypotheses is $\theta_d$. This kind of error is known as a *false positive*, a *type II error* or a *false acceptance* (FA) [Aitken and Taroni 2004; Champod and Meuwly 2000]. If this error occurs, the fact finder will state that the control and recovered materials come from the same source, but they actually come from different sources. Roughly speaking, the fact finder will falsely decide towards the guilt of an innocent person.

Bayes decision rule as stated in Equation 6.8 is optimal if and only if the importance of each type of error is the same. However, this may not be the best criterion in many situations. Imagine, for example, that the consequences of a false positive will be more dramatic than the consequences of a false negative. In order to model the importance of the errors, we define a *cost or penalty* associated to a given type of error, which will be a non-negative real number. The higher the cost, the more important the consequence of a given error. We define $C_{fa}$ as the cost of a single false acceptance error in a decision, and $C_{fr}$ as the cost of a single false rejection error in a decision. In this scenario, the Bayes decision rule aims at minimizing the average Bayes risk, also known as the mean cost [Brümmér and du Preez 2006; Duda et al. 2001; Theodoridis and Koutroumbas 2003], defined as:

$$C_M = P_{fr} \cdot C_{fr} \cdot P(\theta_p) + P_{fa} \cdot C_{fa} \cdot P(\theta_d)$$ (6.11)

In this expression, $P_{fa}$ is the probability of a decision error given that $\theta_d$ is true (a false acceptance error), and $P_{fr}$ is the probability of a decision error when $\theta_p$ is true (a false rejection error). The false acceptance and false rejection probabilities are defined from the likelihoods involved in the decision process in the following way:

$$P_{fr} = \int_{R_d} p(e | \theta_p) \, de$$
$$P_{fa} = \int_{R_p} p(e | \theta_d) \, de$$ (6.12)

where $R_p$ and $R_d$ are the regions where a decision is taken towards $\theta_p$ and $\theta_d$ respectively, according to the value $e = E$. Figure 6.6 illustrates Equation 6.12. The decision regions have to be chosen in order to minimize Equation 6.11. Thus, minimizing Equation 6.11 is equivalent to choosing the regions $R_p$ and $R_d$ so that

\footnote{In this work we will assume that the cost of a right decision will be zero. Although more general approaches may be considered [Brümmér and du Preez 2006; Duda et al. 2001; Theodoridis and Koutroumbas 2003], the
6.2 Bayesian decision theory

Figure 6.6: Probability of false rejection ($P_{fr}$) and false acceptance ($P_{fa}$) and its relationship to $p(E | \theta_p)$ and $p(E | \theta_d)$ respectively.

For a given value of $e = E$,

\[
\begin{align*}
\text{decide } \theta_p &: C_{fr} \cdot p(E | \theta_p) \cdot P(\theta_p) > C_{fa} \cdot p(E | \theta_d) \cdot P(\theta_d) \\
\text{decide } \theta_d &: C_{fr} \cdot p(E | \theta_p) \cdot P(\theta_p) < C_{fa} \cdot p(E | \theta_d) \cdot P(\theta_d)
\end{align*}
\]  

(6.13)

Thus, this easily leads to the following decision rule:

\[
\begin{align*}
\text{decide } \theta_p &: LR > \frac{P(\theta_d) \cdot C_{fa}}{P(\theta_p) \cdot C_{fr}} \\
\text{decide } \theta_d &: LR < \frac{P(\theta_d) \cdot C_{fa}}{P(\theta_p) \cdot C_{fr}}
\end{align*}
\]  

(6.14)

The threshold for which the $LR$ value is compared in order to minimize $C_M$ is known as the Bayes’ threshold [Brümm and du Preez, 2006; Duda et al., 2001], defined as:

\[
\tau_B = \frac{P(\theta_d) \cdot C_{fa}}{P(\theta_p) \cdot C_{fr}}
\]  

(6.15)

and depends on the prior probabilities and the decision costs. The effect of the value of the threshold in the decision boundaries is illustrated in Figure 6.5.

As a conclusion, in order to make optimal decisions in a forensic automatic speaker recognition context according to Bayesian decision theory, the fact finder needs:

1. A $LR$ value from the recovered and control speech. This should be accomplished by the forensic scientist.

---

proposed scenario is suitable for this Thesis.

2Because minimizing the integrand over all regions means minimizing the integral.
2. Setting the Bayes threshold according to the priors and the costs (Equation 6.15). This should be the duty of the fact finder.

Figure 6.7 shows the relationship of all those elements in the decision process in the context of a forensic speaker recognition case.

It is important to note that Bayes decision theory assumes that all the prior probabilities and the likelihoods are being accurately assigned. Under such conditions, computing the LR value would allow the fact finder to take optimal Bayes decisions. However, unavoidable and realistic imperfections in the computation of the LR values will degrade the optimality of the decisions taken by the fact finder, even when Bayes decisions are taken. Next section discusses the accuracy of the computed LR values and its impact on the optimality of Bayes decisions.

6.3. The accuracy of the LR

In this section, we will discuss the evaluation of the accuracy of LR values, describing the concept of calibration of LR values and its importance, as well as previously proposed performance measures for LR values.

6.3.1. Calibration and refinement of LR values

The concept of calibration of posterior probabilities is not new in the statistics literature. In [deGroot and Fienberg 1982] it was introduced in order to evaluate and compare posterior probabilities in the context of weather forecasting. There, posterior probabilities were used as degrees of belief about a given hypothesis (tomorrow it will rain) against its opposite (tomorrow it will not rain). This problem is equivalent to forensic automatic speaker recognition as proposed here, considering posterior probabilities of \( \theta_p \) and \( \theta_d \), computed from prior probabilities and the LR value given by the forensic system.
6.3 The accuracy of the LR

The accuracy of the LR

Figure 6.8: Logarithmic scoring rule. The x-axis represents the posterior probability of \( \theta_p \), which may be viewed as the “forecast” of the fact finder about whether \( \theta_p \) is true considering all the available knowledge in the case.

In deGroot and Fienberg [1982], the accuracy of such a forecaster is assessed by means of strictly proper scoring rules. An example of strictly proper scoring rule, which we will use in this Thesis, is the logarithmic scoring rule. For each value of the evidence \( E \) in a forensic case, the logarithmic scoring rule takes the following values:

\[
\begin{align*}
\theta_p \text{ true} & : - \log_2 (P(\theta_p|E)) \\
\theta_d \text{ true} & : - \log_2 (P(\theta_d|E))
\end{align*}
\] (6.16)

Thus, strictly proper scoring rules may be viewed as loss functions which assign a penalty to a given value of the posterior probability depending on: i) the value of the posterior, and ii) the true value of the hypothesis which actually occurred [Brümmer and du Preez, 2006; deGroot and Fienberg, 1982]. For example, if a probabilistic forecast gives a high posterior probability of raining tomorrow (value of the forecast) and tomorrow it does not rain (true hypothesis), a strictly proper scoring rule will assign a high penalty to the forecast, and vice-versa. The logarithmic scoring rule is illustrated in Figure 6.8.

Strictly proper scoring rules have the following interesting properties:

- Imagine a single forensic automatic speaker recognition trial where, using the prior probability supplied by the fact finder and the LR value yielded by the system, a given posterior probability \( P(\theta_p|E) \) is obtained. We can define the mean value of the proper scoring rule with respect to a reference probability distribution \( Q(\theta_p|E) \), according to deGroot and Fienberg [1982]. This reference probability can be viewed as a desired value of the forecast, to which the actual forecast is compared.

---

2. In deGroot and Fienberg [1982], the reference probability distribution \( Q \) was stated as the actual subjective
For the logarithmic scoring rule, the mean value of the actual forecast with respect to the reference forecast for a single trial will have the following expression:

\[-Q(\theta_p|E) \cdot \log_2(P(\theta_p|E)) - (1 - Q(\theta_p|E)) \cdot \log_2(1 - P(\theta_p|E))\] (6.17)

For a strictly proper scoring rule, such as the logarithmic scoring rule, its mean value for a single trial is minimized if and only if \(P(\theta_p|E) = Q(\theta_p|E)\). In other words, if a forensic automatic speaker recognition system wants to obtain a desired LR value for a given trial in terms of a strictly proper scoring rule, it should yield a LR value which will lead to \(Q(\theta_p|E)\).

In deGroot and Fienberg [1982], the overall measure of goodness of a forecaster is defined as the average value of a strictly proper scoring rule over many different forecasts. For instance, for the logarithmic scoring rule, this mean value would be the logarithmic score (LS):

\[
LS = -\frac{1}{N_p} \sum_{i \in \text{targets}} \log_2 P(\theta_p|e_j) - \frac{1}{N_d} \sum_{j \in \text{nonTargets}} \log_2 P(\theta_d|e_j)
\] (6.18)

Where \(N_p\) and \(N_d\) is the number of target and non-target trials being evaluated. This average value can be viewed as an overall loss. Moreover, it is also demonstrated in deGroot and Fienberg [1982] that such a measure of accuracy can be divided into two components:

1. A calibration loss component, which measures how similar are the forecasts to the frequency of occurrence of \(\theta_p\). Low calibration loss means that for a given range of values of the forecast \(P(\theta_p|E)\) closely around \(x\), the frequency of cases where \(\theta = \theta_p\) tends to be \(x\).

2. A refinement loss component, which measures how sharp or how spread the forecasts are. Roughly speaking, low refinement loss means that if the calibration loss of the forecaster is low, for a given value of the forecast \(P(\theta_p|E)\) the frequency of trials where \(\theta = \theta_p\) is near 0 or 1.

Refinement loss can be seen as discrimination performance as defined in this Thesis in Chapter 5. This is stated e.g. in Brümmer and du Preez [2006]. Moreover, in deGroot and Fienberg [1982], refinement is presented as a measure of the usefulness of the forecaster, regarding the probability of the forecaster. Here, the reference \(Q(\theta_p|E)\) can take other forms. We will discuss it later in this chapter, and we will also choose a reference distribution.

\(^1\)This is easy to check by simply deriving Equation 6.17 with respect to \(P(\theta_p|E)\).
ability to discriminate whether it will rain tomorrow or not. In this work, we will use the term discrimination performance as an equivalent magnitude as refinement loss defined in deGroot and Fienberg [1982].

Both components of the average value of a proper scoring rule over trials, namely discrimination and calibration, have the following interpretation:

- A good refinement means that the scores under analysis have the ability of correctly separating same-source and different-source trials, and therefore it is equivalent as a good discrimination performance.
- The meaning of a good calibration loss is related to the way in which such information is presented to the fact finder. Thus, the aim of calibration is presenting the information contained in the score in order to lead the fact finder to take good decisions in the context of Bayesian decision theory. If the calibration loss is improved, the presentation of the information contained in the score will be better, and the decisions that the fact finder will take on average will be also better [Brümmer and du Preez, 2006; Cohen and Goldszmidt, 2004].

According to deGroot and Fienberg [1982] in this Thesis we propose the averaged value of a strictly proper scoring rule over different trials as the overall measure of accuracy of a given set of LR values. As this measure is empirically obtained, it will depend on the forensic automatic speaker recognition system in use and also on the database and evaluation protocol followed for assessing the LR computation procedure.

6.3.2. Assessing LR accuracy in forensic automatic speaker recognition

6.3.2.1. Classical measures of LR accuracy: Tippett plots

Tippett plots have been classically used for empirical performance assessment [Dessimoz and Champod, 2007; Evett and Buckleton, 1996; Gonzalez-Rodriguez et al., 2006], as in NFI-TNO forensic SRE [van Leeuwen and Bouten, 2004]. Tippett plots are obtained in the same way as the probabilities of false acceptance and false rejection depending on a threshold $\tau$, namely $P_{fa}(\tau)$ and $1 - P_{fr}(\tau)$. However, the meaning of $\tau$ in Tippett plots is not a decision threshold, but the proportion of values for which $LR > \tau$ in the given LR set, conditioned to $\theta_p$ and $\theta_d$ respectively. The relationship between FA/FR plots and Tippett plots is easily seen from Figure 6.9. Important values shown by these curves are the actual distributions of the LR values and the rates of misleading evidence, defined as the proportion of LR values giving support to the wrong hypotheses ($LR > 1$ when $\theta_d$ is true and $LR < 1$ when $\theta_p$ is true) [Dessimoz and Champod, 2007]. In Tippett plots shown in Figure 6.9, the rate of misleading evidence values are highlighted.

\[\text{In fact, some authors represent Tippett plots in the same way as FA/FR rates in Figure 6.9.}\]
6. THE PRESENTATION LEVEL: LR COMPUTATION

![Figure 6.9: Example of Tippett plots showing the actual LR distributions and the rates of misleading evidence when \( \theta_p \) and \( \theta_d \) are respectively true. (a) shows the Tippett plots and (b) their corresponding FA/FR rates plot.

6.3.2.2. Cost-based evaluation of the accuracy of decisions

The decisions taken by an automatic speaker detection system can be evaluated by cost-based techniques, according to the proposed by NIST in their SRE since 1996 [Przybocki et al., 2007]. In Bayes decision theory, the objective is optimizing the mean cost \( C_M \). If the decision costs and prior probabilities are fixed and known, then \( C_M \) is determined using Equation 6.11. There, the values of \( P(\theta_p) \) and \( P(\theta_d) \) for a given threshold \( \tau \) depend on the distribution of FA and FR rates for a given experimental set-up.

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_{fr} )</td>
<td>10</td>
</tr>
<tr>
<td>( C_{fa} )</td>
<td>1</td>
</tr>
<tr>
<td>( P(\theta_p) )</td>
<td>0.01</td>
</tr>
<tr>
<td>( P(\theta_d) )</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 6.1: Values of the costs and the prior probabilities in NIST SRE.

Changing the decision threshold in a speaker detection system leads to different values of \( P_{fr}(\tau) \) and \( P_{fa}(\tau) \), and therefore to different values of \( C_M \). Thus, there is a value of the threshold, namely \( \tau^* \), which leads to a minimum value of the mean cost. According to Brümmer [2004], Brümmer and du Preez [2006], Cohen and Goldszmidt [2004], for a calibrated system the decision threshold determines a pair of \( P_{fr}(\tau^*) \) and \( P_{fa}(\tau^*) \) probabilities which minimize \( C_M \), i.e., \( \tau = \tau^* \). In order to find this optimal threshold, the value of the priors and the costs should be known. For instance, in NIST SRE the priors and costs are given by NIST, and their values...
are shown in Table 6.1. This leads to the $C_{DET}$ mean cost performance metric:

$$C_{DET} = 0.1 \cdot P_{fr} (\tau) + 0.99 \cdot P_{fa} (\tau)$$ (6.19)

As an example, in Figure 6.10(a) the value of $C_M$ for a range of thresholds $\tau$ is represented for different values of the prior and for $C_{fr} = C_{fa} = 1$. It is observed that a minimum value of $C_M$ can be achieved, which is strongly dependent on the prior probabilities. However, for a given forensic case the priors are province of the court and may not be equal among forensic cases. If the priors change, the optimum threshold $\tau^*$ for the original priors and costs will not be optimum anymore for the new priors and costs, as it is observed in Figure 6.10(a). Thus, $C_M$ may dramatically increase because this lack of calibration. Thus, calibration has to be considered for any value of the prior and the costs.

Fortunately, according to Bayesian decision theory, if the automatic speaker recognition system computes LR values, then the optimum threshold for decision making, i.e. the Bayes threshold, is given by $\tau_B = \frac{P(\theta_p)C_{fr}}{P(\theta_p)C_{fr} + P(\theta_d)C_{fa}}$. However, if LR values are not properly computed, an increase in the value of $C_M$ may occur. Figure 6.10 illustrates the consequences of such lack of calibration. Two different systems are shown, and in both cases the value of $C_M$ is represented for a range of thresholds $\tau$ with the constraint that $C_{fr} = C_{fa} = 1$. Figure 6.10(a) shows the system discussed above where LR values have not been computed from the scores, i.e., decisions are taken directly from the scores. Bayes thresholds ($\tau_B$) for each prior are represented as vertical lines. It is clearly observed that the optimality of $C_M$ is very different for different values of the priors. For instance, for $P(\theta_p) = 0.2$ the Bayes threshold almost minimizes $C_M$, but for $P(\theta_p) = 0.5$ the value of $C_M$ dramatically increases far from the optimum. However, Figure 6.10(b) shows a system where LR values have been computed, and it is shown that $\tau_B$ is near from the optimum for all the represented values of $P(\theta_p)$.

Figure 6.10(b) also shows that the optimality of $\tau_B$ will depend on the accuracy of the LR computation process. It is observed that the optimum value of $C_M$ is not exactly in $\tau_B$. This is due to the inaccuracies in the LR computation process. If the value of the LR is not properly computed, then the threshold $\tau_B$ may not be optimal anymore.

### 6.3.2.3. Application-independent assessment of LR values

A solution to the dependency of the priors and costs in order to assess the optimality of the decisions has been proposed in Brümmer and du Preez [2006] for speaker recognition, and since 2006 it has been adopted by NIST in their Speaker Recognition Evaluations (SRE) [NIST]. In Brümmer and du Preez [2006], the term application refers to the values of the priors and the costs in each trial.

In Brümmer and du Preez [2006] a measure of accuracy of a LR values set is proposed,  

\footnote{1In NIST SRE a normalized version of $C_{DET}$ is used, namely Detection Cost Function (DCF). The properties of both metrics are equivalent for our purposes.}
The present level: LR computation

**Figure 6.10:** Value of $C_M$ (Equation 6.11) for different decision thresholds. (a) GMM-SVM-SV system (calibration is not considered) and (b) Logistic regression fused system (calibration is considered). $C_{fr} = C_{fa} = 1$. Bayes thresholds (Equation 6.15) are shown as vertical lines.

namely $C_{llr}$:

$$C_{llr} = \frac{1}{2} N_p \sum_i \log_2 \left( 1 + \frac{1}{LR_i} \right) + \frac{1}{2} N_d \sum_j \log_2 \left( 1 + LR_j \right)$$  \hspace{1cm} (6.20)

where $N_p$ and $N_d$ are respectively the number of target and non-target scores in the evaluation set. The indices $i_p$ and $i_d$ respectively denote summing over the target or the non-target $LR$ set.

In $C_{llr}$ it is easily seen that, for each $LR_i$ value in a given trial, a logarithmic transformation is applied, defined by:

$$\theta_p \text{ true : } \log_2 \left( 1 + \frac{1}{LR_i} \right)$$
$$\theta_d \text{ true : } \log_2 \left( 1 + LR_i \right)$$  \hspace{1cm} (6.21)

It can be seen that $C_{llr}$ is based on the logarithmic scoring rule (Equation 6.16) in the particular case where $P(\theta_p) = P(\theta_d) = 0.5$. Therefore, it presents the attractive properties described for measures of accuracy based on the average value of strictly proper scoring rules over trials.

An important result is derived in Brümmer and du Preez [2006], where it is demonstrated that $C_{llr}$ is the expected decision cost for any value of $C_{fa}$ and $C_{fr}$, averaged over a set of $LR$ values, and assuming $P(\theta_p) = P(\theta_d) = 0.5$. This important result means that minimizing the value of $C_{llr}$ also encourages the obtention of reduced Bayes decision costs for a wide range of values $C_{fa}$ and $C_{fr}$. This result can be generalized to other values of $P(\theta_p)$, and is in fact, as stated in Brümmer and du Preez [2006], the choice of $P(\theta_p) = P(\theta_d) = 0.5$ in order to compute $C_{llr}$ is somewhat arbitrary.

\[ \text{In fact, as stated in Brümmer and du Preez [2006], the choice of } P(\theta_p) = P(\theta_d) = 0.5 \text{ in order to compute } C_{llr} \text{ is somewhat arbitrary.} \]
6.3 The accuracy of the LR

In accordance to other works in the literature [Cohen and Goldszmidt, 2004; Zadrozny and Elkan, 2002], a calibration methodology has been proposed for obtaining LR values by the optimization of $C_{llr}$ with the constrain of preserving the discriminating power of the evaluation score set. If the average of a strictly proper scoring rule is reduced with the constrain of preserving refinement (discriminating power), therefore, by definition, the calibration loss will be reduced. In other words, the DET curve of the evaluation score set must not be changed in this calibration process. As it was discussed previously in Chapter 5, any invertible function does not change the discrimination power of a set of scores, and therefore it does not change the DET curve. Thus, the objective is finding the invertible transformation which optimizes $C_{llr}$ without changing the discriminating power of the system.

Brümmern and du Preez [2006] propose the use of the Pool Adjacent Violators (PAV) algorithm in order to obtain this optimizing transformation. The PAV algorithm implements a technique known as isotonic regression [Zadrozny and Elkan, 2002], and it can be demonstrated that it leads to a transformation which optimizes $C_{llr}$ while preserving the discriminating power of the original score set [Ahuja and Orlin, 2001]. It is important to remark that this PAV-optimization is only achieved if the true answers to the hypotheses are known. Therefore it will not be possible in general to use PAV to optimize $C_{llr}$ on scores of unknown origin. Details about PAV can be found in Brümmern and du Preez [2006]; Zadrozny and Elkan [2002].

In Brümmern and du Preez [2006], PAV is used in order to decompose $C_{llr}$ as follows:

$$C_{llr} = C_{llr}^{min} + C_{llr}^{cal} \tag{6.22}$$

where:

- $C_{llr}^{min}$ Represents the discrimination loss of the system under evaluation. It is obtained by the $C_{llr}$ of the LR values obtained after PAV-optimization. $C_{llr}^{min}$ is the lowest $C_{llr}$ value which a LR set can achieve while preserving the discriminating power of the LR set under evaluation. Therefore, the expected cost due to $C_{llr}^{min}$ is due to non-perfect discriminating power.

- $C_{llr}^{cal}$ Represents the calibration loss of the system under evaluation with respect to the best system preserving discrimination. It is computed as $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$. If the LR values under evaluation converge to the PAV-calibrated LR values, then $C_{llr}^{cal}$ will be reduced.

As an alternative performance measure, the APE plot (Applied Probability of Error) has been also proposed as a performance representation of the LR values computed by the forensic system in a wide range of applications [Brümmern and du Preez, 2006]. The APE plot represents the total probability of error when the Bayes threshold $\tau_B$ is used. This probability is defined as:

$$P_e = P_{fr} (\tau_B) \cdot P (\theta_p) + P_{fa} (\tau_B) \cdot P (\theta_d) \tag{6.23}$$
Figure 6.11 shows an APE plot showing the probability of error of the LR values under evaluation for a wide range of prior probabilities (horizontal axis). The dashed line shows the performance of optimally PAV-calibrated LR values. Thus, the dashed line is a measure of the discriminating power of the system. The dotted curve shows the performance of a neutral system, defined as a system which does nothing, i.e., always outputs a LR value of 1. Therefore, the distance between the dashed and the dotted line is a measure of the calibration loss of the system in terms of $P_e$. As a result, $P_e$ for a system may be high if it is not properly calibrated, even if its discriminating power is good.

![Figure 6.11: Example of APE plot.](image)

In Brümmer and du Preez [2006] it is demonstrated that the integral of the APE curve over the prior is proportional to the value of $C_{llr}$. Therefore it is shown that either improving discrimination or calibration leads to a lower probability of error on average in a Bayesian framework.

APE curves and $C_{llr}$ have been adopted in NIST SRE as official performance measures since the 2006 campaign [Przybocki et al., 2007].

### 6.3.3. Examples of the effects of calibration

The following example illustrates the concepts of discrimination and calibration. Suppose there are two different scores, $e_s$ and $e_d$, which respectively denote a same-source and a different-source comparison. Two different sets of LR values for $e_s$ and $e_d$ are computed. The first set, $\Lambda_1$, has $LR = 5$ for $e_s$ and $LR = 0.5$ for $e_d$. The second set, $\Lambda_2$, has $LR = 5000$ for $e_s$ and $LR = 500$ for $e_d$. For any threshold in $\Lambda_1$ (e.g., $\tau_1 = 2$), there is another threshold in $\Lambda_2$ which will lead to the same false positive and false negative rate (e.g., $\tau_2 = 2000$). Therefore, both sets

\[ P_e \text{ is represented versus the prior log-odds, also known as } \text{logit}(P(\theta_p)) = \log \frac{P(\theta_p)}{P(\theta_d)}. \]
have the same discrimination performance as defined in Chapter 5. However, using $\Lambda_2$ the fact finder will infer posterior probabilities which will lead to erroneous conclusions. For instance, for $e_d$ the LR value in $\Lambda_1$ is 0.5, and the LR value in $\Lambda_2$ is 500. For a prior probability of $P(\theta_p) = 0.5$, $\Lambda_1$ will lead to a posterior probability of $P(\theta_p|E) = 0.333$, from Equation 6.7 which correctly suggests the evidence supports $\theta_d$ for a different-source comparison. However, $\Lambda_2$ will lead to a posterior probability $Pr(\theta_p|E) = 0.998$, from Equation 6.7 which provides misleading evidence in favor of $\theta_p$ for a different-source comparison. Therefore, even when $\Lambda_1$ and $\Lambda_2$ have the same discrimination performance, the lack of calibration in the LR values of $\Lambda_2$ will mislead the fact finder.

Another example about the effects of calibration is shown in this section with an example using synthetic data. Here, three sets of LR values have been synthetically generated for each of the $\theta_p$ and $\theta_d$ hypothesis values simulating three different forensic speaker recognition systems, two of them with the same discriminating ability. Figure 6.12 shows the performance of these synthetic systems (namely System 1, System 2 and System 3) in terms of DET curves, Tippett plots, $C_{llr}$ values and APE curves. DET curves in Figure 6.12(a) show that the discriminating power of System 1 and System 2 are the same in all operating points, outperforming System 3. However, Tippett plots in Figure 6.12(b) show that, although the separation between $\theta_p$ and $\theta_d$ curves is similar in System 1 and System 2, the latter presents a significantly higher rate of misleading evidence for $\theta_d$. Also, posterior probabilities inferred from LR values computed by System 2 and System 3 will lead to important errors because of the high proportion of misleading LR values.

These results are clearly observed in Figure 6.12(c), which presents the same results in the form of $C_{llr}$ values and APE curves. Overall performance is given by $C_{llr}$, split into discrimination loss ($C_{llr}^{\min}$) and calibration loss ($C_{llr} - C_{llr}^{\min}$). It is observed that System 1 and System 2 present the same discrimination performance (same discrimination loss), clearly outperforming System 3. However, $C_{llr}$ values for System 2 and System 3 are quite similar, because of the high calibration loss presented by System 2. On the other hand, the calibration performance of System 1 is the best for all systems.

In order to complete the analysis, APE curves in Figure 6.12(c) show the probability of error for all possible values of prior probabilities if Bayes thresholds are used for making decisions. It is observed that the probability of error dramatically increases when the system is not properly calibrated. Due to this lack of calibration, posteriors inferred using System 2 and System 3 will have a similar probability of error, even when System 2 has a much higher discrimination performance.

\[1\] It can be demonstrated that $\Lambda_1$ and $\Lambda_2$ will have the same DET curve.
6. THE PRESENTATION LEVEL: LR COMPUTATION

![Figure 6.12: DET curves (a), Tippett plots (b) and Callr and APE curves (c) for three simulated systems (System 1, System 2 and System 3). LR values have been randomly generated in order to plot these curves.](image)

6.4. Contribution: information-theoretical evaluation of the accuracy of LR values

Information theory was proposed in the middle of the twentieth century as a framework for measuring and presenting information [Shannon, 1948]. After more than 50 years, the applications of information theory have been remarkable in many fields like physics, probability theory and economics [Cover and Thomas, 2006]. Under this framework, the uncertainty about an unknown variable is quantified by a magnitude called entropy. Additional knowledge about the variables under study will contribute to the reduction of the entropy, and therefore, the information about the unknown variable will rise.

Recently, information theory has been proposed in order to assess the accuracy of automatic speaker recognition systems yielding LR values [Brümmers, 2006; Campbell et al., 2005]. Although these assessment techniques are presented in apparently different forms, they have in essence the same interpretation: the automatic speaker recognition process gives information about whether the control and recovered speech material come from the same speaker.
6.4 Contribution: information-theoretical evaluation of the accuracy of LR values

According to such assessment methodologies, in this Thesis we propose the use of information-theoretical magnitudes for assessing the accuracy of a given set of LR values. The aim of this section is identifying and characterizing the information supplied by the weight of the speech evidence considering the requirements of the so-dubbed *coming paradigm shift* in forensic science [Saks and Koehler, 2005]. The reduction of the uncertainty gives a measure of the expected information about the unknown hypothesis variable $\theta$ that the evidence evaluation delivers to the fact finder, which is modelled in terms of entropy and divergence [Cover and Thomas, 2006]. It will be shown that the proposed measure of accuracy, namely empirical cross-entropy ($ECE$), is in essence the average of the logarithmic scoring rule over different cases, and therefore it agrees with the assessment methodology presented in [deGroot and Fienberg, 1982]. However, our approach makes a clear distinction between the information sources given by the analysis of the evidence, province of the forensic scientist, and the rest of information in the given forensic case, province of the court. A novel performance representation is proposed, namely the $ECE$ plot, which integrates previous approaches and gives a clear and elegant measure of the average reduction of uncertainty supplied by the forensic system. The proposed representation also allows reporting to the court the performance of the forensic system in a clear and simple way, according to the needs of transparency and testability in forensic science.

This section is mainly based on [Ramos and Gonzalez-Rodriguez, 2008].

6.4.1. Information-theoretical evaluation

In this section we will derive an information-theoretical generalization of $C_{\text{llr}}$, namely *Empirical Cross-Entropy* ($ECE$), which measures the accuracy of the LR values for a given value of the prior in terms of average information loss. $ECE$ is essentially a normalized version of other measures proposed in the literature for application-independent evaluation of speaker detection, such as $U_{\log}$ [Brümmen, 2004]. Moreover, another normalized version of $ECE$, namely *normalized cross-entropy* (NCE), has been already proposed in the literature for forensic speaker recognition [Campbell et al., 2005] and in NIST Speech and Rich Transcription evaluations [NIST, 2004].

6.4.1.1. Uncertainty and information

Information theory [Cover and Thomas, 2006; Shannon, 1948] states that the information obtained in an inferential process is determined by the reduction of the entropy, which measures the uncertainty about a given variable in the light of the available knowledge. In our forensic speaker recognition framework, the entropy represents the uncertainty that the fact finder has about the actual value of the hypothesis $\theta = \{\theta_p, \theta_d\}$.

In a given forensic case and before the analysis of the evidence, the uncertainty of the fact finder about the hypotheses is only conditioned to the background information about the case
With this available knowledge, the entropy of the hypothesis, namely prior entropy or entropy of the prior is determined by the following expression [Cover and Thomas, 2006]:

\[ H_P(\theta) = - \sum_{i \in \{p,d\}} P(\theta_i) \log_2 P(\theta_i) \]  

(6.24)

The entropy function is concave with respect to the prior. Its maximum is one (measured in bits), and occurs when \( P(\theta_p) = P(\theta_d) = 0.5 \). Its minimum is zero and occurs when any of the priors equals zero. Thus, entropy is maximum when the uncertainty about the hypotheses is maximum, and entropy is zero when there is certainty about \( \theta \). Figure 6.13 shows the representation of the prior entropy with respect to the value of \( P(\theta_p) \).

![Figure 6.13: Prior entropy with respect to \( P(\theta_p) \).](image)

Once the evidence \( E \) is known and analyzed, a LR value is provided by the forensic system. Then, a posterior probability can be obtained from the prior probability and the LR value using Equation 6.6. In a given forensic case, such LR value may or may not reduce the uncertainty about the hypothesis variable. However, it can be demonstrated [Cover and Thomas, 2006] that the expected value of the entropy of the posterior probability over all possible values of the evidence \( E \) is always smaller than the prior entropy. This expected value is the posterior entropy, computed as [Cover and Thomas, 2006]:

\[ H_P(\theta|E) = - \sum_{i \in \{p,d\}} P(\theta_i) \int_{-\infty}^{\infty} p(e|\theta_i) \log_2 P(\theta_i|e) \, de \]  

(6.25)

where the conditioning evidence value \( E = e \) (here, the value of the score) is integrated over its entire domain.

\(^1\text{Remind that } I \text{ was eliminated from the notation for simplicity, but it conditions every probability value in the case.}\)
6.4 Contribution: information-theoretical evaluation of the accuracy of LR values

The expected information supplied by evidence analysis is illustrated in Figure 6.14. There, it is represented that knowledge about the evidence reduces the expected uncertainty about the hypotheses for all possible values of the evidence [Cover and Thomas, 2006]. However, the computation of Equation 6.25 is usually non-practical, as it requires the knowledge about the likelihoods \( p(e|\theta) \) computed by the system. These likelihoods may not be known in general, e.g., when discriminative LR computation techniques are used as in Brümmer and du Preez [2006]; Campbell et al. [2005]; Gonzalez-Rodriguez et al. [2007b]. Moreover, even when the \( p(e|\theta) \) likelihoods as computed by the forensic system are known, they may not be appropriate for unseen evaluation scores, because of the unavoidable imperfections in the LR computation process especially in forensic conditions (with mismatch between training scores and evaluation scores, scarcity in the training scores, etc.).

A solution to this problem has been proposed in the literature [Brümmer, 2004; Brümmer and du Preez, 2006; NIST, 2004] by comparing the posterior probabilities computed using the forensic system with a reference probability distribution. Thus, the letter \( P \) (for pdfs) will denote probabilities obtained using the forensic system and the letter \( Q \) (for pdfs) will denote reference probabilities. As described before, the reference probability \( Q \) represents a desired value for the posterior, compared to the actual value \( P \) obtained with the system and the prior. This reference probability will be described below in detail.

In order to introduce the reference probability \( Q \), the cross-entropy is defined as:

\[
H_Q \| P (\theta | E) = - \sum_{i \in \{p,d\}} Q(\theta_i) \int_{-\infty}^{\infty} q(e|\theta_i) \log_2 P(\theta_i|e) \, de
\]  

(6.26)

An interpretation of cross-entropy can be given considering that it can be decomposed in the following way:

\[
H_Q \| P (\theta | E) = H_Q (\theta|E) + D_Q \| P (\theta | E)
\]  

(6.27)

where \( D_Q \| P (H|E) \) is the Kullback-Leibler (KL) divergence between the system’s posterior distribution and the reference distribution for all possible values of the evidence [Cover and
Thomas, 2006, defined as:

\[ D_{Q\parallel P}(\theta|E) = \sum_{i \in \{p,d\}} Q(\theta_i) \int_{-\infty}^{\infty} q(e|\theta_i) \log_2 \frac{Q(\theta_i|e)}{P(\theta_i|e)} de \]  

(6.28)

Thus, cross-entropy measures the complementary effect of two different magnitudes:

- \( H_Q(\theta|E) \), the posterior entropy of the reference, which measures the uncertainty about the hypotheses if the reference probability distribution is used for computing posteriors.
- \( D_{Q\parallel P}(\theta|E) \), the deviation of the system’s posterior \( P \) from the reference posterior \( Q \).

This is an additional information loss, because it was expected that the system computed \( Q \), not \( P \).

6.4.1.2. Proposed measure of accuracy: empirical cross-entropy (ECE)

The computation of the cross entropy via the integral in Equation 6.26 may be tedious if possible. However, an empirical procedure is given here in order to approximate its value. Given a target and a non-target set of LR values from forensic testing, we can obtain two target and non-target sets of posterior probabilities using Equation 6.2, assuming that the prior probabilities are known. Therefore, we can average the expectations in Equation 6.26, supposing the law of the large numbers holds, obtaining:

\[
ECE = - \frac{Q(\theta_p)}{N_p} \sum_{i \in \text{targets}} \log_2 P(\theta_p|e_j) \\
- \frac{Q(\theta_d)}{N_d} \sum_{j \in \text{nonTargets}} \log_2 P(\theta_d|e_j) 
\]  

(6.29)

where:

\[ H_{Q\parallel P}(\theta|E) \simeq ECE \]  

(6.30)

This value will be our evaluation objective, namely \textit{empirical cross-entropy} (ECE), which is equivalent to the already proposed NCE [Campbell et al., 2005; NIST, 2004] and \( U_{\log} \) Brümmer, 2004.

An important result is highlighted here: \( ECE \) is the (prior-weighted) average value of the logarithmic scoring rule over the whole set of trials in the defined experimental set-up. As a consequence, it complies with the method for accuracy assessment proposed in deGroot and Fienberg 1982 and described before in this chapter. Thus, the value of \( ECE \) considers the discrimination performance of the LR set being assessed, but also its calibration. This is a main motivation for defining the accuracy of a LR values set in terms of \( ECE \).
6.4 Contribution: information-theoretical evaluation of the accuracy of LR values

6.4.1.3. The role of the prior in ECE

The role of the prior probability is highlighted here, since the posterior probability is dependent on the prior probability \( P(\theta_p) \) and the LR value:

\[
P(\theta_p | E) = \frac{LR \cdot \frac{P(\theta_p)}{P(\theta_d)}}{1 + LR \cdot \frac{P(\theta_p)}{P(\theta_d)}}
\]  

(6.31)

Then, ECE can then be expressed as:

\[
ECE = \frac{Q(\theta_p)}{N_p} \sum_{i \in \text{targets}} \log_2 \left( 1 + \frac{1}{LR_i \cdot \frac{P(\theta_p)}{P(\theta_d)}} \right)
+ \frac{Q(\theta_d)}{N_d} \sum_{j \in \text{nonTargets}} \log_2 \left( 1 + LR_j \cdot \frac{P(\theta_p)}{P(\theta_d)} \right)
\]  

(6.32)

Thus, ECE is prior-dependent, and it is not possible in general for the forensic scientist to compute its value for a given particular case, because the prior probabilities in such a case are province of the fact finder. However, it is possible to compute and represent ECE for a set of values of the prior in order to represent ECE in court. Then, the fact finder can compute the ECE value for a particular prior in a particular case. Thus, the exact value of the prior is irrelevant for the forensic scientist and the prior is taken as a parameter. This prior-dependent representation will be the proposed approach for presenting ECE in court, as it will be shown later.

6.4.2. ECE and information loss

The contribution to the information loss modelled by ECE is illustrated in Figure 6.15, where the average information loss measured by cross-entropy is shown in terms of its decomposition (Equation 6.27). As remarked before, the prior is taken as a parameter, and therefore \( H_P(\theta) = H_Q(\theta) \). Therefore, from Equation 6.32 it is straightforward that ECE is independent of the

Figure 6.15: Information change due to evidence analysis and assignment of reference probabilities. The cross-entropy consists of the posterior entropy of the reference (inner ellipse, uncertainty) plus the divergence between the reference and the posterior probability of the system (red darker area in outer ellipse, information loss).
reference probability $Q$. Thus, the selection of $Q$ is only constrained by Equation 6.27. This has the following interpretation: for a fixed value of $ECE$, changing the reference $Q$ implies that:

- $H_Q(\theta|E)$ increases (decreases) and
- $D_{Q\|P}(\theta|E)$ decreases (increases)

in order to keep $ECE$ constant. This is illustrated in Figure 6.15: the ellipse representing cross-entropy has always the same size. However, the inner small grey ellipse representing posterior entropy of $Q$ may increase or decrease depending on the choice of the reference $Q$.

### 6.4.2.1. Choosing a reference $Q$ for intuitive interpretation

The selection of the reference probability $Q$ is constrained, because Equation 6.27 must hold. Therefore, the reference $Q$ may be carefully selected. Moreover, in order to interpret the results in court, simplicity and clarity should be the objective. Therefore, in this chapter we propose a selection of the reference probability distribution $Q$ which has an intuitive interpretation in the context of a forensic case. It may be derived as follows: the aim of every forensic case is finding the true value of the hypothesis variable $\theta$. If the fact finder already knew such true values, he would always obtain the following posterior probabilities:

$$Q(\theta_p|E) = 1, \text{ if } \theta_p \text{ is true}$$

$$Q(\theta_d|E) = 0, \text{ if } \theta_d \text{ is true}$$

The posterior distribution in Equation 6.33 will be referred to as the oracle distribution. If this oracle distribution is selected as a reference $Q$, then the entropy of the reference posterior $Q$ is zero ($H_Q(\theta|E) = 0$) and therefore the $ECE$ becomes the $KL$–divergence $D_{Q\|P}(\theta|E)$ of the posterior distribution of the system with respect to the oracle posterior.

The choice of such a reference posterior has an attractive and simple interpretation: the higher the $ECE$ value, the more the average information the fact finder needs over cases in order to know the true value of the hypotheses. If the LR values of the forensic automatic speaker recognition system are misleading to the fact finder, then the $ECE$ will grow, and more information on average will be needed in order to know the true values of the hypotheses.

### 6.4.3. The $ECE$ Plot

In this Thesis we propose a representation of $ECE$ as a function of $P(\theta_p)$ in a so-called $ECE$ plot. For each prior probability in a partition of the $[0, 1]$ range, posterior probabilities are computed using the LR values for the evaluation set, computed by the forensic automatic speaker recognition system, and Equation 6.6. The value of $ECE$ (Equation 6.32) is then represented as a function of $P(\theta_p)$.

Figure 6.16(a) shows an example of $ECE$ plot for a sample system. The solid curve is the $ECE$ (average information loss) of the LR values computed by the system. The higher this
6.4 Contribution: information-theoretical evaluation of the accuracy of LR values

Figure 6.16: Comparison of the ECE plot (a) to APE plots and $C_{llr}$ (b). ATVS SVM-SV system in NIST SRE 2006. A high calibration loss is observed, as no LR computation technique has been performed on the scores.

ECE curve, the higher the information needed in order to know the true hypotheses on average over cases, and therefore the worse the system. Two other systems are also represented for comparison:

- On the one hand, the dashed curve represents the calibrated system, which optimizes ECE preserving discrimination. As proposed in Brümmer and du Preez [2006], the calibrated system is obtained from the forensic system using the Pool Adjacent Violators (PAV) algorithm (see Brümmer and du Preez [2006] for details).

- On the other hand, the dotted curve represents the performance of a system always delivering $LR = 1$, referred to as a neutral system. The posterior in this case is the prior, which is independent of the system. Thus, according to Equation 6.26 the cross-entropy of such a system is simply the entropy of the prior probability, given by Equation 6.24 and represented in Figure 6.13.

In an ECE plot, both comparative systems play an important role.

Calibrated system. It represents the component of the ECE due to the non-perfect discriminating power of the LR set under analysis, because the component of ECE due to calibration has been minimized. It also represents the best accuracy for any set of LR values having the same DET plot of the original LR set.

Neutral system. If the ECE value of the forensic system is greater than the entropy of the neutral system, then the forensic system will loss more information on average than basing the decisions only on the prior information, i.e., not using the forensic system. In the range of prior probabilities where this happens, the forensic system should not be used for evidence analysis.
The \(ECE\) plot is easy to interpret if we choose the oracle reference. Imagine a case in court where a control and a recovered sample are presented as evidence. The fact finder asks for the forensic evidence evaluation of the speech samples. Suppose that the fact finder establishes a given value for \(P(\theta_p)\) before the analysis of the evidence. Thus, the \(ECE\) value in the plot at the given value of \(P(\theta_p)\) is the average information (over forensic cases) that we need in order to know the true value of the hypothesis for the given prior.

6.4.3.1. Information and invertible transformations

As has been stated above, a reduction in the uncertainty of \(\theta\) gives information. This is quantified by the mutual information between the evidence \(E\) and the hypothesis \(\theta\) [Cover and Thomas, 2006]:

\[
I(\theta; E) = H(\theta) - H(\theta|E)
\] (6.34)

Mutual information must be non-negative, and the higher \(I(\theta; E)\), the more information the evidence \(E\) gives to the hypothesis \(\theta\) (and vice-versa, because \(I(\theta; E) = I(E; \theta)\)). Another interesting property of mutual information is due to the data processing inequality. Let \(\phi(\cdot)\) any arbitrary function of the evidence \(E\). Then, it can be demonstrated that:

\[
I(\theta; E) \geq I(\theta; \phi(E))
\] (6.35)

In other words, processing the evidence never increases the information which is given to the hypothesis. The demonstration of data processing inequality can be found in [Cover and Thomas, 2006].

Now consider an invertible transformation of the evidence, namely \(\psi(\cdot)\). Data processing inequality states that:

\[
I(\theta; E) \geq I(\theta; \psi(E))
\] (6.36)

However, it is also stated that:

\[
I(\theta; \psi(E)) \geq I(\theta; \psi^{-1}(\psi(E))) = I(\theta; E)
\] (6.37)

Equations (6.36) and (6.37) simultaneously hold if and only if:

\[
I(\theta; E) = I(\theta; \psi(E))
\] (6.38)

Therefore, applying invertible transformations to \(E\) preserves the information which the evidence gives to the hypotheses. This is a fundamental result, which has been previously sketched in the literature [Brümer and du Preez, 2006]. By this property, applying invertible transformations to the score of an automatic speaker recognition system never reduces the information in the score about the hypothesis \(\theta\). \(LR\) computation methods which are invertible transformations of the score will be referred to as information-preserving. As discrimination is also preserved if an
invertible transformation is applied to the scores as it was shown in Chapter 5, then invertible transformations are discrimination- and information-preserving.

6.4.4. Comparison to other performance measures

$ECE$ and $C_{lr}$ are closely related. From Equations (6.20) and (6.32) it is straightforward that

$$C_{lr} = ECE|_{P(\theta_p) = 0.5}$$

Thus, $C_{lr}$ is a value which summarizes $ECE$. The interpretation of $C_{lr}$ in terms of information is now straightforward. It measures the average information needed by the fact finder in order to know the true values of the hypotheses when the prior uncertainty is maximum (i.e., if $H(\theta) = 1$). Moreover, following the derivation of $C_{lr}$ in Brümmer and du Preez [2006], $ECE$ at a given prior represents the expected cost of making decisions over a wide range of decision costs, whose value at $P(\theta_p) = 0.5$ is $C_{lr}$.

Regarding APE plots, it can be demonstrated [Brümmer and du Preez, 2006] that the integral of the APE plot over the prior is proportional to $C_{lr}$. Therefore, reducing the error rate for $\tau_B$ also reduces the value of $ECE$ at $P(\theta_p) = 0.5$. A comparison between an $ECE$ plot and an APE curve is shown in Figure 6.16. It is shown that the $ECE$ value gives similar intuition about the accuracy (discrimination plus calibration) of the system as the APE plots. Also, it is clearly shown that the value of $C_{lr}$ is the value of $ECE$ at $P(\theta_p) = 0.5$.

6.4.5. Experimental example

In order to show the adequacy of the proposed information-theoretical assessment methodology, we present experimental results using two ATVS-UAM systems and its fusion, namely the GMM and the GMM-SVM-SV systems. NIST SRE 2006 protocol was followed in order to conduct the tests. Details about the experimental set-up can be found in Chapter 3. The two systems have been fused via logistic regression [Brümmer et al., 2007], which not only fuses the scores coming from the individual systems, but it also tends to calibrate them. Logistic regression will be deeply described later in this chapter, and has been performed using the FoCal toolkit.

In Figure 6.17 the $ECE$ plots are shown for the individual systems and for their fusion via logistic regression. Figures 6.17(a,b) shows that the $ECE$ values are not satisfactory for the individual systems. Actually, if a fact finder assumes that he will receive a $LR$ value from the system and the system did not take into account calibration, the decisions of the fact finder may be dramatically far from the optimum, which is represented by a growth of $ECE$. Hence, in the case of the individual systems the $ECE$ is far from its calibrated value, as it can be seen from the difference between the dashed and solid curves.

---

Footnotes:

1. The value of $ECE$ at $P(\theta_p) = 0.5$ has been highlighted with a dashed line in the $ECE$ plot in order to easily find $C_{lr}$.

6. THE PRESENTATION LEVEL: LR COMPUTATION

Figure 6.17: ECE plots for the individual systems based on SVM-SV (a) and GMM (b), and for the fused system using logistic regression (c).

Figure 6.17(c) shows the ECE plot for the fused system. It is observed that the ECE value is smaller than for the individual systems, which is justified by the calibrating transformation applied by logistic regression. This improvement is observed for all priors. Also, the difference between the dashed and solid curves is small, which means a small information loss due to calibration.

6.5. Transforming scores into accurate LR values

Once a measure of accuracy has been stated by means of the average of strictly proper scoring rules over trials (such as ECE), in this section we will describe how to transform scores computed by the forensic automatic speaker recognition system into accurate LR values.

The starting point for LR computation in the proposed methodology is the similarity score, computed at the discrimination level. The aim of the presentation level is yielding an accurate LR value without degrading the discrimination performance of the original score set. In this Thesis we will classify such LR computation methods in two main classes:

1. **Pure calibration.** With these techniques, the discrimination performance of the input scores is unchanged. This is accomplished by the use of an invertible transformation, which preserves the discrimination performance of the scores. Such techniques may be also referred to as information-preserving, because the mutual information shared by the hypotheses and the score remains the same.

2. **Hybrid discrimination-calibration.** The aim of such techniques is finding a score-to-LR transformation which not only improves the calibration, but also the discrimination performance. Thus, this kind of LR computation techniques are not limited to an invertible transformation. Of course, the objective is also further improving the discrimination performance of the input scores.

Figure 6.18 illustrates these two LR computation schemes. The main advantages of pure
6.5 Transforming scores into accurate LR values

Figure 6.18: Proposed classification of LR computation techniques. Pure LR computation does not change discriminating power, whereas hybrid discrimination-calibration should not degrade discriminating power.

calibration techniques against hybrid discrimination-calibration rely in the fact that the discrimination of the scores is unchanged by the presentation stage. That may be important in several frameworks where the LR computation process is performed under unfavorable conditions. On the other hand, the main drawback of pure calibration against hybrid discrimination-calibration is the invertibility constraint, which does not allow further improvement of the discrimination performance of the output LR values.

The main motivation of the proposed classification relies on the score-based architecture of most automatic speaker recognition systems, where discrimination performance is usually the main measure of goodness. However, it is important to remark that any LR computation technique based on case-specific training data will generally lead to a non-invertible score-to-log (LR) transformation if scores are pooled among cases, even if the case-by-case calibration scheme is invertible. This may happen if LR computation considers case-specific populations or if within-source variability modelling depends on the specific suspect involved in the case, as it happens in typical forensic casework. Here we classify the transformations among invertible and non-invertible if it is assumed that the same transformation is applied to all scores from evidence evaluation. Although it is difficult to have such situation in practical casework, it may be assumed in order to assess system accuracy in controlled conditions.

Next sections in this chapter will introduce several LR computation techniques found in the literature according to the proposed classification. Then, an experimental comparison will be performed with the NIST SRE 2006 benchmark.
6.5.1. Pure score calibration techniques

6.5.1.1. Logistic regression

Logistic regression is a well-known pattern recognition technique widely used for many problems including fusion [Brümm et al., 2007; Pigeon et al., 2000] and more recently calibration [Brümm and du Preez, 2006; Gonzalez-Rodriguez et al., 2007b]. The aim of logistic regression is obtaining an affine transformation (i.e., shifting and scaling) of an input dataset in order to optimize an objective function. Let \( E = \{E_1, E_2, \ldots, E_K\} \) be a set of scores from \( K \) different automatic speaker recognition systems. The affine transformation performed by the logistic regression model can be defined as:

\[
f_{lr} = \log \left( \frac{O(\theta_p|E)}{1 + O(\theta_p)} \right) = a_0 + a_1 \cdot E_1 + a_2 \cdot E_2 + \ldots + a_K \cdot E_K
\] (6.40)

Using Bayes theorem in odds form (Equation 6.6) this expression allows the computation of the logarithm of the LR value for a given prior.

\[
\log(LR) = a_0 + a_1 \cdot E_1 + a_2 \cdot E_2 + \ldots + a_K \cdot E_K - \log \left( \frac{O(\theta_p)}{1 + O(\theta_p)} \right)
\] (6.41)

It is easily seen from Equation 6.7 that such model leads to the following expression for \( P(\theta_p|E) \), namely the logistic regression model:

\[
P(\theta_p|E) = \frac{1}{1 + e^{-f_{lr}}} = \frac{1}{1 + e^{-\log(LR) - \log(O(\theta_p))}}
\] (6.42)

As it can be seen, the logistic regression transformation from score values to posterior log-odds is invertible unless it is a constant function. The weighting terms \( \{a_0, a_1, a_2, \ldots, a_K\} \) can be obtained from a set of training data with optimization procedures found in the literature. For a given value of the prior odds \( O(\theta_p) \) we may define \( f_{lr}^t = a_0 + \sum_{j=1}^{K} a_j E_j^t \) as the value obtained for a given set \( E^t = \{E_1^t, \ldots, E_K^t\} \) of target scores from the automatic speaker recognition systems. On the other hand, let define \( f_{lr}^{nt} = a_0 + \sum_{j=1}^{K} a_j E_j^{nt} \) the value obtained for a given set \( E^{nt} = \{E_1^{nt}, \ldots, E_K^{nt}\} \) of non-target scores from the \( K \) automatic speaker recognition systems. Logistic regression computes the \( \{a_0, a_1, a_2, \ldots, a_K\} \) coefficients by making \( P(\theta_p|E) \) as close as possible to 1 for target trials and to 0 for non-target trials, contrained to the logistic regression model (Equation 6.42). It can be derived [Brümm et al., 2007; Pigeon et al., 2000] that such optimization leads to the following objective to minimize:

\[
C_{wlr} = P(\theta_p) \cdot \frac{1}{N_t} \sum_{i=1}^{N_t} \log_2 \left( 1 + e^{-f_{lr,i}^t} \right) + P(\theta_d) \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} \log_2 \left( 1 + e^{f_{lr,i}^{nt}} \right) = ECE
\] (6.43)

---

1 In this Thesis we have used the FoCal toolkit for training logistic regression models (http://niko.brummer.googlepages.com).
where \( N_t \) is the number of \( f^{t}_{lr} \) values and \( N_{nt} \) is the number of \( f^{nt}_{lr} \) values, both in the training set. As it is highlighted in Equation 6.33, the optimization objective in logistic regression is precisely the Empirical Cross-Entropy (ECE) of the training score set. For a given value of \( P(\theta_p) \) and assuming a logistic regression model, ECE is convex with respect to the weighting terms \( \{a_0, a_1, a_2, \ldots, a_K\} \), and therefore it has a global minimum [Brümmer et al., 2007].

As logistic regression optimizes ECE, it can be used for calibration as well as for fusion. If the number of systems \( K \) is more than one, we will be fusing the input scores and mapping them to a single value of the posterior log-odds. As an additional effect, because of the objective function optimization, the output of such a fusion will tend to be calibrated. On the other hand, if \( K = 1 \) then the input score is transformed by an affine mapping which will tend to give a calibrated output.

Therefore, for a given value of the prior probabilities, but using Equation 6.41, posterior log-odds are mapped again into log (LR) values which will also tend to be calibrated. If the prior probabilities are known, the value of \( P(\theta_p) \) can be set. If the prior probabilities are unknown, the log (LR) value can be obtained for an arbitrary value of the prior, and it will tend to be calibrated for any value of the prior. A typical choice for this prior may be \( P(\theta_p) = 0.5 \), and therefore logistic regression will optimize \( C_{llr} \).

### 6.5.1.2. Other pure score calibration techniques

Several LR computation techniques using invertible transformations have been proposed in the literature. In [Brümmer et al., 2007], a nonlinear version of logistic regression is proposed, namely \( s-cal \). Its motivation relies in the fact that LR values strongly supporting the wrong hypothesis (strongly misleading evidence) will have a very big penalty in terms of the logarithmic scoring rule. On the other hand, LR values supporting the correct hypothesis will have a low penalty. Therefore, it seems reasonable to limit strongly supporting LR values in order to avoid high penalties. Thus, \( s-cal \) applies a trained, strictly monotonic saturation to the output values. The resulting function is a strictly monotonically rising sigmoid, which is an invertible function.

Another approach to pure calibration has been proposed by the use of the Pool Adjacent Violators (PAV) algorithm [Brümmer and du Preez, 2006]. The PAV algorithm transforms a set of scores into a set of calibrated LR values. However, it is only possible to apply an optimal PAV transformation if the true hypotheses for each score in the set are known. Nevertheless, as suggested in [vanLeeuwen and Brümmer, 2007], a PAV transformation can be trained on a set of scores for which the true value of the hypotheses are known and then apply the trained transformation to scores for which the hypothesis value is unknown.

This approach has a drawback, since the PAV transformation for a given score set is step-wise, and therefore it will not be invertible in general for any score set. Thus, a straight use of PAV leads to a non-invertible transformation. However, several smoothing techniques can be applied to PAV in order to keep it monotonically rising. For instance, adding an infinitely small slope to PAV will lead to an invertible transformation. Interpolating with linear, quadratic or splines approaches are also possible smoothing schemes.
Figure 6.19: (a) FA/FR rate distributions of a real ATVS-UAM speaker recognition system in NIST SRE 2006. (b) Comparison of pure-calibration score to log (LR) transformations.

Figure 6.19 shows a comparison of the score-to-log (LR) transformations for the algorithms described here. It can be seen that all transformations are monotonically rising, and therefore invertible. That means that the discriminating power of the original scores is preserved. Moreover, for the s-cal and PAV cases, the range of the output LR values is limited.

6.5.2. Hybrid score discrimination-calibration

6.5.2.1. Generative likelihood modelling

LR computation in forensic automatic speaker recognition has been classically performed by the use of generative techniques modelling the hypotheses-conditional distribution of the evidence scores $E$. This is the approach already presented in Meuwly [2001], and has been followed in subsequent works in the literature Alexander [2005], Drygajłk [2007], Gonzalez-Rodriguez et al., 2006, Ramos-Castro et al., 2006b. Under this approach, the objective is assigning the likelihoods $p(e|\theta_p)$ and $p(e|\theta_d)$ respectively to the target and non-target scores, in order to compute the LR value.

Assigning $p(e|\theta_p)$ and $p(e|\theta_d)$ implies the selection of a proper model. Several different approaches have been found in the literature, being the most popular the use of single Gaussian distributions Botti et al., 2004, Ramos-Castro et al., 2005b and Kernel Density Functions (KDF) Alexander, 2005, Gonzalez-Rodriguez et al., 2005a, Meuwly, 2001. Figure 6.20 shows several likelihood modelling strategies over the same set of scores, with their corresponding score-to-log (LR) transformations. As it is observed, the resulting transformation will be non-monotonic in general, and therefore non-invertible. However, there are some particular cases where the resulting transformation is invertible, such as the case of single Gaussian modelling with equal variances.
6.5 Transforming scores into accurate LR values

Figure 6.20: Comparison of generative hybrid discrimination-calibration LR computation techniques. Left figure shows the pdfs assigned to the data (the scores used are the same as in Figure 6.19). In (a) and (b), Gaussian distributions are assumed, having equal variances in (b). In (c) and (d), Kernel Density Functions with different smoothing parameters are used.
6.5.2.2. Multilayer perceptrons

The use of neural networks has been recently proposed for the LR computation problem. In [Campbell et al., 2005], a multilayer perceptron (MLP) [Duda et al., 2001; Theodoridis and Koutroumbas, 2003] is used in order to obtain a transformation which optimizes normalized cross-entropy (NCE), a normalized version of ECE. Therefore, the MLP approach presented in [Campbell et al., 2005] can be viewed as a non-linear version of logistic regression. Thus, the MLP approach encourages the obtention of calibrated log (LR) values. Another advantage of MLP is the possibility of easily including other kind of meta-information such as the quality of the input signal, speech utterance durations, etc. In [Campbell et al., 2005] it is demonstrated that such information contributes to a reduction in the value of ECE.

Since the MLP approach is in general non-linear and there are no constraints about the invertibility of the final score-to-log (LR) transformation, the resulting mapping will be generally non-invertible, and therefore it is needed to check the impact of the transformation in the discriminating power of the output LR values.

6.5.3. Contribution: generative suspect-adapted LR computation

As shown in the literature [Gonzalez-Rodriguez et al., 2006], the accuracy of LR computation is especially affected by small sample size effects due to control speech data scarcity coming from a given suspect. Several techniques have been proposed in the literature in order to compensate for data scarcity. The framework proposed in [Botti et al., 2004] assumes that an accurate model of the within-source distribution for a given suspect can be obtained using target scores from different individuals in the same conditions. Therefore, the within-source distribution $p(e|\theta_p)$ is modelled from scores coming to different speakers. However, it has been shown that, even in the same conditions, the target scores coming from different speakers may present different distributions [Doddington et al., 1998]. Therefore, the accuracy in within-source estimation may be improved by exploiting suspect-specific scores.

In this Thesis we propose a novel technique, namely suspect-adapted LR computation, which achieves robustness by exploiting speaker-independent information and suspect specificities using an adaptive approach. The proposed technique is based on generative modelling of within- and between-source distribution, and therefore it is not expected to preserve discrimination performance of the input scores, but to improve it due to the exploitation of the specificities of the speech coming from a given speaker.

6.5.3.1. MAP adaptation for suspect-adapted within-source variation modelling

Our strategy is based on the adaptation of the speaker-independent target score distribution to the suspect target scores via MAP (Maximum A Posteriori) inference [Gauvain and Lee, 1994]. Let $E_G = \{E_{G1}, \ldots, E_{GN}\}$ be a set of global target scores computed using speech from speakers other than the suspect. Let $E_S = \{E_{S1}, \ldots, E_{SM}\}$ be a set of suspect target scores.
6.5 Transforming scores into accurate LR values

Global target scores in $X_G$

Suspect target scores in $X_S$

Global pdf $g_G(x)$

Suspect pdf $g_S(x)$

Adapted within-source $g_A(x)$

Figure 6.21: Example of MAP adaptation using synthetic scores.

scores obtained from the control speech obtained from the suspect involved in the forensic case. First, using Maximum Likelihood and assuming Gaussian distributions, we estimate the pdfs $p_G(e) = N(\mu_G, \sigma_G)$ and $p_S(e) = N(\mu_S, \sigma_S)$ from $E_G$ and $E_S$ respectively, where:

- $p_G(e)$ represents the variability of target scores between speakers.
- $p_S(e)$ represents the variability of target scores coming from the suspect.

Assuming a single Gaussian model, we assign the distribution in the numerator of the $LR$ Equation 6.3 as follows: $p(e | \theta_p) \equiv p_A(e) = N(\mu_A, \sigma_A)$. We compute the within-source distribution parameters using MAP adaptation as follows [Gauvain and Lee, 1994]:

\[
\begin{align*}
\mu_A &= \alpha \mu_S + (1 - \alpha) \mu_G \\
\sigma^2_A &= \alpha (\sigma^2_S + \mu^2_S) + (1 - \alpha) (\sigma^2_G + \mu^2_G) - \mu^2_A
\end{align*}
\]

(6.44)

The adaptation coefficient $\alpha$ is defined as

\[
\alpha = \frac{M}{M + r}
\]

(6.45)

and depends on: 1) the number of suspect scores $M$ and 2) a fixed relevance factor $r$. From Equations 6.44 and 6.45 it is observed that when $M$ is small, the algorithm gives more importance to global data $E_G$. As more suspect scores are available, the adapted within-source distribution will be more adjusted to the suspect data $E_S$. Note that if $r = 0$ then $p_A(e) = p_S(e)$. On the other hand, if $r \to \infty$ then $p_A(e) \to p_G(e)$ and the resulting within-source will be speaker-independent as in Botti et al. 2004. Figure 6.21 illustrates this technique for $r = 1$ from suspect and global target scores generated synthetically.
6.5.3.2. Experiments

In order to validate the proposed suspect-adapted LR computation technique, experiments have been performed using the 8conv4w-1conv4w test condition in the evaluation protocol proposed in NIST 2005 SRE. We carry out our experiments using the ATVS GMM system with feature warping compensation technique. See Chapter 3 for details about the experimental set-up.

In order to obtain each suspect’s target score set \( E_S \), we have selected all the target scores for each speaker from the whole score set in the evaluation, except the score used as evidence in each LR computation. Thus, there will be a variable number of within-source scores for each speaker. We have only selected suspect vs. questioned speech comparisons having more than four suspect target scores, i.e., \( M \geq 5 \). A total number of 10,618 trials have been performed in this sub-condition. All the process has been carried out in a gender-dependent way, and no cross-gender trials have been performed.

Before the evaluation, a development set consisting of the 8conv4w-1conv4w condition for NIST 2004 SRE database was selected. Trials performed using this development set follow the NIST 2004 SRE protocol. The global target score set \( E_G \) consists of all the target scores in this development set. As the database used in NIST 2004 was also a different subcorpus of the Mixer database, \( E_G \) is supposed to accurately represent the global variability of all suspect scores in the test set. The population used for LR computation consists of respectively 224 female and 170 male speaker GMM models in the development set.

![Figure 6.22: \( C_{tr} \) for suspect-adapted LR computation with different amount of suspect scores \( M \) and different relevance factors \( r \) for the selected subset of the 8conv4w-1conv4w condition in NIST 2005 SRE.](image)

In order to simulate a lack in the suspect data in the selected subset of the 8c-1c condition of NIST 2005 SRE, we randomly select subsets of \( M \) scores from the total number of suspect target scores in each LR computation, which is done for different values of \( M \). Thus, we evaluate the effect of a lack of target suspect scores fixing the rest of conditions. Performance of the system
for different relevance factors $r$ (Equation 6.45) and different number of suspect target scores $M$ is shown in Figure 6.22. We have computed the $C_{llr}$ for different values of $M$ and $r$. As a result, it can be observed that, for $M \leq 4$, the system performance tends to its optimum value for $r = 1$. Thus, the proposed speaker-adapted technique outperforms speaker-dependent ($r = 0$) and speaker-independent ($r \to \infty$) within-source models. Also, if $M \geq 5$ the best results are obtained for $r = 0$, although similar performance is obtained for values of $r$ close to 1. In other words, the proposed technique performs properly for any amount of suspect scores, not only in data scarcity situations. It is also seen that the system performance is quite stable from $r = 0.5$ to $r = 2$.

Tippett plots in Figure 6.23 show the performance of the system using MAP adaptation for different values of $M$ and for $r = 1$. The claimed robustness can be observed in Figure 6.23 as $M$ decreases, the distribution of LR values is similar, especially in the sense of misleading LR values, i.e., $LR > 1$ values when $\theta_d$ is true and $LR < 1$ values under $\theta_p$.

### 6.5.4. Experimental comparison of LR computation methods

In this section, several LR computation methods previously described will be compared in an experimental way. The NIST SRE 2006 database and protocol will be used in order to present these comparative results. Scores will be generated with the ATVS-UAM GMM system with feature mapping. From such scores, LR values will be computed using linear logistic regression, generative Gaussian distributions and the proposed generative Gaussian suspect-adapted method. Scores for training the LR computation methods have been generated using the ATVS-UAM GMM system and the NIST SRE 2005 database and protocol. See Chapter \[9\]
6. THE PRESENTATION LEVEL: LR COMPUTATION

Figure 6.24: DET plots showing discrimination performance for the ATVS-UAM GMM system for different LR computation methods.

for details about systems and experimental protocols.

In Figure 6.24 DET plots are shown, which present the discrimination performance of the compared LR computation techniques. It is observed that the scores directly computed by the automatic speaker recognition system have exactly the same discrimination performance as the LR values computed using logistic regression, because logistic regression is invertible. Moreover, Gaussian (suspect-independent) modelling presents very similar performance. This is because the score-to-log (LR) parabolic transformation is monotonic in a wide range of the domain where the input scores lay. However, suspect-adapted Gaussian modelling is not invertible, as a different mapping is applied to each suspect. Therefore, the suspect-adapted technique changes the discrimination performance, which in our case is improved for the LR values as it can be observed in Figure 6.24.

Figure 6.25 shows the distribution of LR values in the form of Tippett plots for the scores and LR values obtained. In Figure 6.25a, it can be seen that the magnitude of the strength of misleading evidence (LR values supporting the wrong hypothesis) is high for the scores when \( \theta_d \) is true, which suggests a bad accuracy if the scores are going to be used directly as \( \log (LR) \) values. On the other hand, after LR computation the rate of misleading evidence is quite limited for all the presented techniques (Figures 6.25b-d). However the Tippett plots do not allow us to easily conclude which technique is more accurate. Moreover, it is hard to distinguish if the systems degrade due to discrimination or to calibration.

Figure 6.26 shows the APE plots and \( C_{llr} \) values for the presented techniques. It is observed in APE plots that, using Bayes thresholds, the total probability of error \( P_e \) is much higher for the scores before LR computation. That will lead to a higher mean cost for many different values of \( C_{fa} \) and \( C_{fr} \), whose expected value is \( C_{llr} \). After LR computation, the probability

126
6.5 Transforming scores into accurate LR values

Figure 6.25: Tippett plots showing distribution of scores and \( \log_{10}(LR) \) values for the ATVS-UAM GMM system for different LR computation methods.
of error decreases, and $C_{ltr}$ reduces substantially, which means that the overall performance when making Bayes decisions is better on average over a wide range value of prior odds. The values of $P_e$ for the calibrated system (given by the PAV algorithm) and $C_{ltr}^{min}$ show that the discrimination performance of the scores is not significantly changed by the LR computation process, and even though slightly improved by the suspect-adapted technique. This effect was also observed in DET plots before (Figure 6.24).

Finally, Figure 6.27 shows the performance for the presented LR computation techniques in the form of ECE plots. It can be seen that for values of $\log_{10}(O(\theta_p))$ between $-1$ and $+1$, the cross-entropy of the scores (Figure 6.27, solid line) is higher than for all LR values computed using the presented techniques. It is also observed that such ECE value is far from the calibrated system after PAV. That means that if the scores are directly used as $\log(LR)$ values they will lead to a loss of information due to miscalibration, on average over trials. However, ECE (solid curve) dramatically reduces after LR computation, and therefore the fact finder will need less average information in order to know the true value of $\theta$. Furthermore, the ECE value for the system is far better than the neutral system (dotted curves), which justifies the use of the system for evidence evaluation. Finally, for all the LR computation schemes presented, the ECE of the calibrated system (dashed curve) is near the ECE of the LR values computed by the system (solid curve), which demonstrates the good calibration performance of these techniques.
6.5 Transforming scores into accurate LR values

Figure 6.27: ECE plots showing the discrimination and calibration performance of the ATVS-UAM GMM system using information-theoretical magnitudes, for different LR computation methods.
6.6. **LR computation for non-score-based systems**

The proposed methodology for automatic forensic speaker recognition considers the similarity score as the output of the discrimination stage because the wide majority of speaker recognition systems in the literature are score-based as it is justified in Chapter 4. However, a natural approach for LR computation works directly on the features, without the computation of an intermediate, non-calibrated similarity score. This has been the methodology classically followed in forensic science for continuous features [Aitken and Taroni, 2004; Lindley, 1977], except in the case of score-based biometric systems.

Computing LR values directly from features has been also proposed for phonetic-acoustic forensic speaker recognition [Gonzalez-Rodriguez et al., 2007b; Rose, 2006a]. Moreover, such a scheme has been proposed by Jiang and Deng [2001] and Vogt and Sridharan [2004] for GMM-based speaker recognition systems. In this case, the LR values are computed directly from the features following a Bayesian inferential process similar as the one proposed by Aitken and Lucy [2004].

For such LR computation methods, it seems that the proposed methodology cannot be applied, as there is not a score computation process between the control and recovered speech samples and the LR value. However, a simple variation in the proposed hierarchical methodology allows transparent and testable forensic automatic speaker recognition with such kind of LR computation methods. In this case the discrimination and presentation levels converge into one single abstraction level, where score computation and optimization of the calibration loss of LR values are performed at once. Figure 6.28 shows this combined discrimination-presentation level proposed for non-score-based systems. The performance evaluation of LR values yielded at this level will have to consider both discrimination and calibration performance, e.g., by means of ECE plots as proposed in this Thesis.

6.7. **Chapter summary and conclusions**

In this chapter we have described the presentation abstraction level in the proposed hierarchical methodology for forensic automatic speaker recognition. The objective at this level is yielding a LR value as a meaningful degree of the support of the evidence to whether the control and the recovered speech material come from the same source or not. Inputs and outputs of this level have been detailed, as well as their relationship to the rest of levels in the hierarchy.

First, the LR methodology for evidence evaluation has been described, highlighting the role and meaning of the LR value. Bayes decision theory has been then addressed as a way of making optimal decisions. The importance of properly computing LR values has been remarked, and a definition of the accuracy of the LR value has been given as a measure of the goodness of the evidence evaluation process. Such a definition has been adopted from the statistics literature by means of the use of strictly proper scoring rules. The important concept of calibration of LR values has been derived from such approach. Then, several ways of measuring the accuracy of LR
values found in the literature have been reviewed, and a novel information-theoretical approach to the assessment of LR values has been proposed as a contribution. Thus, in this Thesis we propose the use of empirical cross-entropy ($ECE$) and $ECE$ plots in order to transparently represent the accuracy of LR values.

After that, we have reviewed several techniques found in the literature in order to transform scores coming from a forensic speaker recognition system into accurate LR values. A novel approach has been also proposed, namely suspect-adapted generative LR computation. The proposed technique improves discrimination and calibration of LR values by the exploitation of the specificities of the suspect under analysis. The proposed scheme has been experimentally compared to other LR computation methods proposed in the literature, following the NIST SRE 2006 protocol. Several performance measures are used in order to present results, including the proposed $ECE$ plots.

The chapter ends with an alternative configuration for the hierarchical methodology proposed here in order to include non-score-based LR computation methods common in other disciplines in forensic science. This is important for the generality of the approach, as classical procedures for LR computation in forensic science obtain LR values directly from continuous features, without an intermediate score computation process.

Original contributions in this chapter are the definition of accuracy of LR values in the context of forensic identification; the information-theoretical measure of accuracy $ECE$; the representation using $ECE$ plots; the classification of LR computation methods presented; and the suspect-adapted generative LR computation technique.
Chapter 7

The forensic level: forensic reporting
and presentation in court

This chapter presents the last abstraction level in the proposed hierarchical methodology for forensic automatic speaker recognition: the forensic level. This level is essential, since it is the interface with the court. Hence, considerations about reporting and presenting results with forensic automatic speaker recognition systems are covered in this level. The objective is unifying the results from evidence analysis and forensic testing with the demands of the court and the needs derived from the coming paradigm shift in forensic identification.

As it was justified in Chapter 4, DNA as golden standard should be followed as a model in order to present scientifically sound results in court. Therefore, the weight of the forensic evidence should be presented in a data-driven and probabilistic way, and the way of its obtention should be clearly reported. Moreover, the report should reflect that the whole process is transparent and repeatable, allowing the fact finder to scrutinize each step in the evidence evaluation and the assessment of the techniques used.

The forensic level belongs to the proposed hierarchical methodology for speaker recognition, as it has been defined in Chapter 4. In Figure 7.1 the inputs and outputs of this level are detailed. On the one hand, this level receives the following inputs:

- **LR value (from presentation level).** The LR value computed by the forensic automatic speaker recognition system represents the degree of support to any of the hypothesis involved in the trial, which allows taking decisions in a Bayesian framework.

- **Discrimination and calibration performance (from presentation level)** of the LR values is essential at the forensic level, since it determines the accuracy of the LR system.

- **Background information about the case.** The knowledge about the circumstances of the case may be essential in order to present assessment results relevant to the court, and also for population selection issues. It strongly determines which database to use in each case.
Speech databases. Reporting of speech data used for assessment is essential in order to present results according to the requirements of the court, regarding the evaluation of the evidence and also the results of assessment tests.

Court demands. In order to generate the report about the case, it is fundamental to consider the demands of the court, which determines the whole evidence evaluation and interpretation process.

On the other hand, the forensic level yields the following outputs to the court:

- Forensic report. This is the information the fact finder will receive about the evidence evaluation process and the procedures used for it. It is essential that this report copes with the needs of the court, considering the coming paradigm shift in forensic science.

- Testing results. The accuracy of the methods used for evidence evaluation should be clearly stated, according to the Daubert admissibility rules and other similar criteria. In particular, the intuitive interpretation of the accuracy of the system and a clear determination of the adequacy of the procedures in use should be key points for reporting assessment results.

This chapter is organized as follows. First, we give some guidelines in order to report the weight of the speech evidence and its accuracy in accordance to the coming paradigm shift in forensic science. Then, we give several examples about how to explain the assessment results expressed in the form of the information-theoretical quantities proposed in this Thesis. We
illustrate such examples with simulated cases using two different forensic automatic speaker recognition experimental set-ups: i) the ATVS-UAM GMM system using NIST SRE 2006 protocol (Chapter 3), and ii) a real forensic speaker recognition system used by Spanish Guardia Civil, with a database collected from field casework recordings, namely BDRA. We then illustrate how the proposed assessment techniques can be generalized to forensic disciplines other than automatic speaker recognition, by the presentation of an experimental example for glass and paint evidence. Finally, conclusions are drawn.

Original contributions in this chapter include the methodology for presenting results from assessment experiments based on ECE plots, as proposed in this Thesis; the experimental example using systems and data from real casework; and the extension of this methodology to other forensic disciplines such as glass and paint evidence.

7.1. Reporting to the court with forensic automatic speaker recognition systems

7.1.1. The coming paradigm shift and forensic automatic speaker recognition reporting

The forensic level is mainly based on the adaptation of the results obtained in the presentation level to the needs of the court considering the coming paradigm shift in forensic identification science. Regarding forensic automatic speaker recognition, here we give some guidelines in order to converge to the uprising needs in forensic science, which have been identified in Chapter 4:

- **Transparency.** Scrutinizing and comparing the procedures for evidence evaluation and interpretation is a key issue in a forensic case. The methods used for forensic evaluation should be adapted to a possible need of the court in this sense. Thus, they must be transparent, which implies that procedures, statements and arguments in a forensic report should be based on well-grounded scientific disciplines. In forensic automatic speaker recognition, signal processing and pattern recognition are the main disciplines involved, which are widely accepted and reviewed by the scientific community. Also, in order to clearly report LR values to the court, all evidence evaluation procedures should be also based on statistics. The use of this disciplines together with a data-driven approach as proposed in this Thesis allow fact finders to scrutinize the whole process.

- **Testability.** Forensic reports should include any assessment test demanded by the court. Thus, database availability is essential, but it is also important to know which database to use for testing and evidence evaluation [Champod et al., 2004]. Moreover, it is essential that the report includes a clear description of the testing procedures used, as well as the database used for assessing the techniques. In forensic automatic speaker recognition, appropriate available methodologies are those from main benchmarks in the area, such as
NIST and NFI-TNO evaluations [Przybocki et al., 2007; van Leeuwen et al., 2006] as has been used in this Dissertation.

- **Accuracy.** A forensic report should define accuracy in an intuitive and straightforward way, previous to the assessment of the methods in use. Moreover, the adequacy of the technique should be also stated in relation to the given definition of accuracy. In forensic automatic speaker recognition, the definition and the measures of accuracy described and proposed in this Thesis (ECE plots, $C_{lit}$, APE plots) are appropriate, especially because their intuitive interpretation.

- **Common procedures.** A forensic report will be easier to interpret if a common methodology for presenting evidence and forensic tests is used. Fortunately, in automatic speaker recognition, widely accepted NIST and NFI-TNO benchmarks [Przybocki et al., 2007; van Leeuwen et al., 2006] represent a common procedure for benchmark tests. Moreover the use of LR methodology following DNA helps fact finders to correctly interpret the weight of the evidence. Also, the presentation of results by means of methodologies accepted in the automatic speaker recognition community significantly helps the convergence to a transparent framework.

### 7.1.2. Speech variability and forensic automatic speaker recognition reporting

As it has been highlighted in Chapters 2 and 5, the variability in the conditions of the speech for different sessions is one of the challenges in state-of-the-art automatic speaker recognition technology. Moreover, the forensic environment is far from ideal, and it is typical in casework to face recordings presenting different transmission channels, environmental noise, disguised voices, variable speaker’s mood, etc. Session variability compensation techniques have been significantly improved in the last years, but it is still a challenging problem in some particular environments such as typical forensic casework. Due to this high variability in the speech signal between sessions, several considerations are in order when reporting to the court using forensic automatic speaker recognition systems.

**Mismatching conditions in evidence evaluation.** Mismatch in evidence evaluation when comparing control and recovered speech segments should be carefully handled in order to avoid degradations in the accuracy of the LR values computed. General guidelines for handling mismatch in evidence evaluation have been discussed through this Thesis, as well as in Alexander [2005]; Gonzalez-Rodriguez et al. [2006]. However, it is important to remark this issue in forensic reports not only by clearly stating the conditions of all the speech involved, but also by presenting assessment experiments which give an idea of the degradation of the accuracy due to mismatching conditions.

**Mismatching conditions accuracy assessment.** Most widely extended benchmarks in automatic speaker recognition, such as NIST SRE or NFI-TNO evaluation, consider speech
quality and length similar to those used in many cases involving lawful telephone interceptions. However, a major question to be addressed remains in every case: are the conditions of the assessment equivalent to those of the recovered questioned recording, suspect control speech and relevant population in the case? Possible mismatch between known databases used for assessment and casework speech data remains a major problem and an active research topic. The influence of any mismatch on system performance must be addressed in forensic reports by the submission of comparative experiments showing system performance in conditions similar to those used in each case, in order to preserve transparency in the assessment process.

**Population conditions.** The populations for generating non-target scores for $LR$ computation should be selected in a case-by-case basis, depending on the conditions of the recovered speech segment, the conditions of the control recording from the suspect and the circumstances of the case. It is out of the scope of this thesis to discuss how to select a proper population for forensic reporting, and there are several works in the literature addressing this issue [Alexander, 2005; Champod *et al.*, 2004; Meuwly, 2001]. However, it is important to note the need of clearly stating the population conditions in the forensic report. Unlike DNA and other disciplines, in forensic speaker recognition the characteristics of the population does not depend only on the characteristics of the individuals whose speech was recorded, but also on the environmental conditions of the recordings, which have to be clearly stated in the report.

### 7.2. Presenting results in court using forensic automatic speaker recognition systems

In this section, two experimental examples are reported in order to illustrate the process of evidence evaluation and its presentation in court. The objective here is simply illustrating how to state the weight of the evidence, the accuracy of the system and the possible issues about admissibility and adequacy of the proposed method for its use in a forensic case.

We report two different simulated cases using two different systems. First, the ATVS-UAM GMM system presented in Section 3.2.3 is used in order to generate scores for the simulated case. Second, we use a database, protocol and system used by the Spanish Dirección General de la Guardia Civil (DGGC) used in real forensic casework.

#### 7.2.1. Experimental example: the use of the ATVS forensic speaker recognition system in a simulated case

Suppose a scenario, adapted from a real case, where the prosecutor presents a piece of evidence consisting of an incriminating questioned recording obtained from the wing of a prison.

---

1We strongly thank Major José Juan Lucena, from the Forensic Acoustics department of Spanish Guardia Civil, for giving permission to use the scores derived from their systems and databases.
containing 11 inmates. A suspect is appointed from police investigations, one of the inmates in the wing, and some speech recordings are obtained from him. Considering only this background information, the fact finder may assign a prior probability $P(\theta_p) = \frac{1}{11}$ (1 over 11) that $\theta_p$ is true (the suspect is the source of the questioned speech). The court gives the forensic speech scientist both recordings, and information relative to the suspect and the circumstances of the case. The fact finder also insists the scientist’s analytical technique must comply with Daubert-like rules.

The forensic scientist, taking into account all those elements, uses the ATVS-UAM GMM system to generate scores, and suspect-adapted $LR$ computation in order to evaluate the weight of the evidence. The performance of such evidence evaluation technique is shown in Figure 7.2 for NIST SRE 2006. In order to generate non-target scores for training the $LR$ computation techniques, the forensic scientist selects speech samples coming from a population of individuals. Speech recorded are obtained from the population considering the circumstances of the case, according to the guidelines described in the literature, such as [Alexander 2005; Gonzalez-Rodriguez et al. 2006; Meuwly 2001]. The selection of all databases involved in the $LR$ computation process should be clearly documented in the forensic report, as well as the justification for such selection.

Moreover, considering the admissibility requirements of Daubert rules, the forensic scientist decides to include in the report results of forensic testing taking into account the circumstances and conditions of the analyzed recordings. Possibly among other performance measures (DET plots, Tippett plots, etc.), the scientist includes the $ECE$ plot of the forensic test in order to explain the fact finder the average information given by the system in the inferential process, as proposed in this Thesis. This average information will be measured in $bits$, as measure of binary information. It is important to include in the report that a bit contains the information needed to know the value of the hypothesis variable $\theta$ in a single case, in order to help interpretation of $ECE$.

As stated before, it is important that the forensic scientist will clearly specify the environmental conditions of the speech used in the forensic test. If conditions are different from those in the case, convenient warnings should be also included.

If the fact finder so desires, the scientist may explain in court the accuracy of the evidence evaluation method used by the forensic scientist. She decides to use the $ECE$ plot for that, as a measure of how the average information would be improved over many forensic cases by the use of the forensic system. For the ATVS-UAM GMM system with suspect-adapted $LR$ computation, presented in Figure 7.2, the argument of the scientist should be as follows:

- Before knowing the weight of the evidence, and given that the prior odds have been set to $O(\theta_p) = 0.1$ (i.e., the $-1$ value on the $ECE$ plot x-axis), the $ECE$ plot shows that we need 0.45 bits of information on average in order to know the true value of the hypothesis averaged over cases like this one (dotted curve of Figure 7.2(d) at $O(\theta_p) = 0.1$).

1 For the presented example, no other information is assumed to be present in the forensic case. However, in a real case the background information may include either more circumstantial information or other evidence sources.
7.2 Presenting results in court using forensic automatic speaker recognition systems

Figure 7.2: Performance of the LR values computed in the simulated case using the ATVS-UAM GMM system and suspect-adapted evidence evaluation. DET plots (a), Tippett plots (b), APE plots and $C_{llr}$ (c) and finally ECE plots (d) are shown for the scores of the ATVS-UAM GMM system and after suspect-adapted LR computation.
After analyzing the weight of the evidence, more information has been obtained, and we will need only 0.18 bits on average in order to know the true value of the hypothesis averaged over cases like this one (solid curve of Figure 7.2(d) at \( O(\theta_p) = 0.1 \)).

If we had used the calibrated system we would have needed 0.14 bits on average in order to know the true value of the hypothesis (dashed curve of Figure 7.2(d) at \( O(\theta_p) = 0.1 \)). However, it has to be clear that this calibrated results are not feasible in practice, because the forensic scientist needs to know information about the true answers of the hypotheses in order to obtain this calibrated system.

It is also important to remark to the court that, if only this information-theoretical assessment is considered, it would be worth to admit the presented evidence evaluation for the given forensic case. The reason is that, at the prior odds of 0.1, the value of \( ECE \) (solid) is smaller than the entropy of the neutral system (dotted curve). In other words, on average over forensic cases, the evidence evaluation process gives information about the value of the hypothesis (\( \theta_p \) or \( \theta_d \)), because the difference between the dotted curve and the solid curve is positive. The higher this difference, the more information will be supplied to the fact finder on average over different forensic cases.

### 7.2.2. Experimental example: simulated case with real forensic data and systems

In this section, a reporting example is provided using systems and databases from real forensic casework. The objective is showing the performance and illustrating some issues about database collection in real forensic work. The system presented in this example is being used in the laboratories of the Spanish Guardia Civil. It is based on LR computation algorithms proposed in [Gonzalez-Rodriguez et al. 2006](#), and therefore presents robustness to scarcity in speech data coming from the suspect.

Guardia Civil usually deals in casework with wire-tappings coming from cellular networks, which are then recorded in magnetic tapes. In order to handle such cases, the Department of Forensic Acoustics in Dirección General de la Guardia Civil (DGGC) has made a significant effort for collecting a database in the same conditions of such traces, namely cellular wire-tappings recorded in magnetic tape. The result has been the obtention of a database called BDRA (Base de Datos de Registros Acústicos, Acoustic Registers DataBase), registered in the Spanish Agency of Data Protection, and aiming at scientific research\textsuperscript{1}. The subcorpus used in this Thesis consists of 65 speakers, identified by police investigations, and recorded in analog magnetic tape from Spanish GSM networks and in some cases analog cellular networks. The speakers are male individuals speaking in native Spanish from Spain, and they come from 14 different cases between 1992 and 2004.

\textsuperscript{1}See Orden Ministerial INT-3764-04, BOE 277 (Spanish Official Bulletin) for details.
From the 65 speakers in the sub-corpus of BDRA, 35 have been used as relevant population for \( LR \) computation, and the rest have been used to generate \( LR \) values. Speaker models have been built using 2 minutes of speech after manually silence removal. 10 test segments have been extracted from the database for each of the 30 speakers, having lengths between 10 and 15 seconds per utterance. All the possible combinations have been used for generating \( LR \) values, leading to \( 30 \times 10 = 300 \) target and \( 30 \times 29 \times 10 = 8,700 \) non-target \( LR \) values for the experimental set-up. The performance of the DGGC system using such a database for forensic testing is illustrated in Figure 7.3.

Suppose again the example in Section 7.2.1: an incriminating, wire-tapped recovered recording from a stolen GSM telephone in the wing of a prison. Suppose that this recording has been recovered by Guardia Civil from a GSM network and it has been stored in magnetic tape. Now imagine that a suspect is appointed, and he admits to be the author of several non-incriminating control recordings obtained in the same way by Guardia Civil over one month period.

Under these conditions, a comparison between the control and the recovered speech materials will lead to a \( LR \) value to be reported in court. Moreover, BDRA database will result extremely useful in order to select a population with the appropriate recording conditions in order to compute the \( LR \) value, assuming that such database is representative of the population defined by the circumstances of the case. Furthermore, assessment results presented in Figure 7.3 will appropriately represent the performance of the forensic automatic speaker recognition system for the control and recovered material being analyzed. Therefore, the forensic scientist will be able to reason in court in the following way.

- Before knowing the weight of the evidence, and given that the prior odds have been set to \( O(\theta_p) = 0.1 \) (i.e., the \(-1\) value on the \( ECE \) plot x-axis), the \( ECE \) plot shows that we need 0.45 bits of information on average in order to know the true value of the hypothesis over cases like this one (dotted curve of Figure 7.3(d) at \( O(\theta_p) = 0.1 \)).

- After analyzing the weight of the evidence, more information has been obtained, and we will need only 0.18 bits on average in order to know the true value of the hypothesis over cases like this one (solid curve of Figure 7.3(d) at \( O(\theta_p) = 0.1 \)).

- If we had used the calibrated system, which preserves the discrimination capabilities of the technique in use, we would have need 0.17 bits on average in order to know the true value of the hypothesis (dashed curve of Figure 7.3(d) at \( O(\theta_p) = 0.1 \)). However, it has to be clear that this calibrated results are not feasible in practice, because the forensic scientist needs to know information about the true answers of the hypotheses in order to obtain this calibrated system.

- We remark that the conditions of the incriminating recovered speech recording, the control speech sample for which the suspect has recognized to be the author and the assessment results presented are similar, i.e., wire-tappings from GSM networks recorded in magnetic
Figure 7.3: Performance of the DGGC system in the simulated case using the BDRA database. DET plots (a), Tippett plots (b), APE plots and \( C_{llr} \) (c) and finally ECE plots (d) are shown for the LR values delivered by the forensic automatic speaker recognition system.
7.3 Generalization to other forensic disciplines

In this section, we present several results which demonstrate that the proposed methodology can be extended to other forensic disciplines. In particular, we present ECE plots in order to assess the accuracy of the evaluation of glass and paint evidence.

7.3.1. The LR meaning and accuracy across forensic disciplines

In this Thesis, we have always particularized LR computation for score-based automatic speaker recognition systems. Moreover, an alternative configuration of the proposed methodology has been proposed in order to cope with non-score based LR computation methods which directly compute LR values from features or measurements extracted from the speech signal. However, this methodology can be extended to forensic disciplines other than automatic speaker recognition. For instance, in Gonzalez-Rodriguez et al. [2007], a similar methodology than the one proposed in this Thesis is used both for traditional (phonetic-acoustic) and automatic forensic speaker recognition, integrating both approaches. In fact, LR-based analysis of the evidence has been proposed as a common framework of convergence for evaluating the evidence for all forensic disciplines [Evett, 1998].

We can conclude that the interpretation of the LR value in a Bayesian context is independent among disciplines [Aitken and Taron, 2004]. The LR value is always a degree of support to a prior opinion about the hypotheses involved in the case (expressed as prior odds). Thus, the LR value as a numerical expression of the weight of the forensic evidence will have the same meaning for different forensic disciplines.

Moreover, the evaluation techniques for measuring the accuracy of the LR computation methods, such as ECE plots, are also independent of which LR-based forensic discipline is being used. This is because such assessment techniques are only dependent on the LR values to be assessed, no matter where these LR values come from. Therefore, although LR values have the same meaning across forensic disciplines, they can have very different accuracies for different LR computation methods and for different forensic disciplines.

For instance, a forensic automatic speaker recognition system will possibly present a much higher ECE value than a DNA profiling system in forensic tests representing real casework conditions. In that case, the accuracy of forensic automatic speaker recognition would be significantly worse than for DNA profiling. However, the interpretation of a LR value as a degree
of support in a given forensic case is always the same, no matter if it comes from a forensic automatic speaker recognition system or from a DNA test.

7.3.2. Experimental example: glass and paint evidence

In order to illustrate the adequacy of the methodology proposed in this Thesis for different forensic disciplines, we present some assessment results from experiments performed using glass and paint evidence. This section is mainly based in [Ramos et al. 2007].

7.3.2.1. LR computation technique

Measurements from glass and paint objects typically consist of a vector of concentrations of several discriminating elements in the glass case (e.g., aluminum, magnesium, sodium, etc.) or organic compounds in the paint case (e.g., styrene, methacrylate, 1,6-diisocyanatohexane, etc.). Thus, the comparison of two glass or paint samples based on such measurements is in essence a multivariate problem. In order to obtain a LR value from multivariate data, the method proposed by [Aitken and Lucy 2004] can be used. However, databases of chemical glass and paint profiles are usually scarce, and therefore the probabilities computed using multivariate analysis can be inaccurate if the dimensionality of the problem rises (an effect which has been dubbed curse of dimensionality [Duda et al. 2001]). Therefore, by [Aitken et al. 2007] a dimensionality reduction procedure based on graphical models is proposed in order to decompose a high-dimensional problem into several low-dimensional sub-problems, which will be later factorized. Such two-level model for LR computation has been used for glass and paint experiments. Details can be found in [Aitken et al. 2007; Ramos et al. 2007].

7.3.2.2. Glass evidence experimental framework

Four glass fragments (Figure 7.4), with surfaces as smooth and flat as possible, collected from each of 270 glass objects (104 building windows, 63 car windows, 26 bulbs, 16 headlamps, 57 containers, 6 glasses) were analyzed by Scanning Electron Microscope coupled with X-ray Detector (SEM-EDX), allowing the determination of concentrations of oxygen (O), sodium (Na), magnesium (Mg), aluminum (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe). From the measurements on these elements, seven independent variables were derived by taking log10 of the measurements on the seven other elements normalized to the oxygen measurements, leading to \( Na', Mg', Al', Si', K', Ca' \) and \( Fe' \) variables to be analyzed.

Two hundred objects from the glass database were used as background data for glass evidence in order to compute statistics for evidence evaluation using the techniques described in [Aitken et al. 2007]. The remaining 70 objects were used in order to compute LR values and validate

---

1Permission to publish assessment results from glass and paint measurement databases has been kindly given by Grzegorz Zadora, from the Institute of Forensic Research in Krakow, Poland (http://www.ies.krakow.pl). The author also thanks him for providing the experimental protocols for comparison.
the approach. That yielded a total of 70 same-source comparisons and \(0.5 \times 70 \times 69 = 2415\) different-source comparisons.

**Figure 7.4:** A microscopic image of glass fragment.

### 7.3.2.3. Paint evidence experimental framework

The paint database was composed of 20 samples of acrylic topcoat paints (Figure 7.5) that were indistinguishable on the base of their infrared spectra and elemental composition. From pyrolytic gas chromatography, seven variables were obtained from the analysis of the following components: 2-hydroxyethylmethacrylate (M2E - abbreviation used in this Thesis); \(\alpha\)-methylstyrene (MST); toluene (TOL); butylacrylate (BMA); 2-hydroxypropylmethacrylate (M2P); methylmethacrylate (MMA); 1,6-diisocyanatehexane (I16).

For paint evidence, a division between background data for training and testing was not practical due to the lack of available data (20 objects). Therefore, the whole paint database was used either as background data and as testing data. Thus, there were 20 same-source and \(0.5 \times 20 \times 19 = 190\) different-source comparisons for paint evidence. Although this approach is too optimistic, it illustrates the proposed assessment methodology. Future work considers extending the paint database in order to have a more realistic experimental protocol for paint evidence.

### 7.3.3. Results

#### 7.3.3.1. Information-theoretical analysis

In this section, ECE plots are presented with the purpose of showing the average information loss of the LR values computed by the forensic scientist over different forensic cases. In Figures 7.6(a-c) the performances of three of the bivariate LR values computed for the glass model are
Figure 7.5: A cross-section of car paint observed under an optical microscope (magnification 40x)

Figure 7.6: ECE plots for glass evidence. In (a-c) examples of three sets of bivariate LR values are shown, namely \((Na', Si')\), \((K', Na')\) and \((Ca', Na')\). In (d) the ECE plot is shown for the LR for the proposed technique based on graphical models.

shown as an example. The ECE of the calibrated LR values (dashed curve) is fairly similar (ranging from 0.4 to 0.6 for \(P(\theta_p) = 0.5\), except for a set of variables \((Ca', Na')\) where its value reaches 0.9). That means that the discrimination performance is similar for all the presented bivariate LR methods. On the other hand, the ECE of the LR values computed by univariate
### 7.3 Generalization to other forensic disciplines

#### Figure 7.7: ECE plots for paint evidence. Examples of two sets of univariate and one set of bivariate LR values: TOL (a), MST (b) and (MMA,M2P) (c). In (d), the result for the LR values computed using graphical models is shown.

and bivariate problems (solid curve) varies significantly. Moreover, for some bivariate variables like \((Ca', Na')\) (Figure 7.6c) the value of the ECE is high and comparable to the neutral LR set (LR is always 1). Despite the discrimination capabilities of \((Ca', Na')\), the average information of the LR values computed by the forensic scientist for \((Ca', Na')\) will be almost the same as the average information of a neutral expert (LR = 1 always).

Figure 7.6d shows the performance of the LR computation technique based on graphical models for glass evidence. It shows that the ECE of the calibrated LR values (dashed curve) is better for the graphical model than for the univariate and bivariate models (ECE of 0.23 for the graphical model for \(P(\theta_p) = 0.5\)). Thus, LR values computed with the proposed graphical model present a much better discrimination performance than the univariate and bivariate LR values. However, the ECE of the LR values actually computed by the forensic scientist is quite high because of miscalibration, which limits the average information delivered by the scientist’s LR values. This may be due to the effect of some of the bivariate LR values which have a high value of ECE, like \((Ca', Na')\) (Figure 7.6c). The difference in ECE between the scientist’s LR values and the calibrated LR values indicates that while the scientist’s LR values provide good discrimination (as shown in Aitken et al. 2007 by means of false positives and false negatives) there is still a considerable room for improvement in order to reduce calibration loss.
The results of a similar analysis for three univariate \( LR \) values for paint evidence are shown in Figures 7.7a, 7.7b, and 7.7c. The performance of the calibrated \( LR \) values (dashed curve) in this case is excellent (ranging from 0.1 to 0.4 for \( P(\theta_p) = 0.5 \)). Moreover, the difference between the \( ECE \) of the calibrated \( LR \) values and the scientist’s \( LR \) values is much lower than for the glass evidence case, which means an amount of average information delivered by the scientist’s \( LR \) values closer to the optimal calibration case. Also, the \( ECE \) plot for the graphical model for paint evidence is shown in Figure 7.7d. Again, the loss of information due to discrimination (dashed curve) is excellent (\( ECE = 0.05 \) for \( P(\theta_p) = 0.5 \)). However, although the performance of the forensic scientist’s \( LR \) set (solid curve) is fairly good for prior probabilities higher than 0.5, the dashed and solid curves are still separated. Therefore an improvement of the calibration would lead to an even better accuracy. This is especially needed for \( P(\theta_p) < 0.1 \), where the scientist’s \( LR \) values are even more inaccurate on average than the neutral \( LR \) set.

7.4. Chapter summary and conclusions

The forensic abstraction level in the hierarchical methodology for forensic automatic speaker recognition proposed in this Thesis has been described. At this level, court demands and forensic science upcoming needs are taken into account aiming at reporting the weight of the evidence and its assessment to the fact finder, and also presenting it to the court in a transparent way.

The issue about forensic reporting to the court using automatic speaker recognition systems has been analyzed in the light of the identified requirements derived from the coming paradigm shift in forensic science (Chapter 4), and several guidelines for forensic reporting have been highlighted for speech evidence. Then, the presentation of the weight of the evidence and the assessment of its potential accuracy in court has been illustrated by two examples, one of them using the systems, databases and protocols used by Spanish Guardia Civil in real casework. Important topics such as the impact of mismatching conditions, the transparency in the limitations of forensic testing and some guidelines for minimum system requirements have been illustrated by means of the information-theoretical assessment framework proposed in this Thesis.

Finally, the chapter presents an experimental example which illustrates the adequacy of the proposed methodology for other forensic disciplines such as glass and paint analysis. It is highlighted that, whereas the \( LR \) values computed for different techniques across multiple forensic disciplines may have a significantly variable accuracy, its meaning remains the same [Aitken and Taroni, 2004; Evett, 1998]. Therefore, the methodology and the techniques presented in this Thesis may be used for any other \( LR \)-based forensic discipline.

The review of the identified reporting requirements; the guidelines for presentation of forensic results in court using \( ECE \); and the considerations about the generalization of the proposed methodology to other forensic disciplines; are all original contributions.
Chapter 8

Conclusions and future work

This Thesis has addressed the problem of the evaluation of the forensic evidence using automatic speaker recognition systems. A hierarchical methodology for the use of automatic speaker recognition systems in forensic cases has been proposed. This methodology supposes a powerful tool for practitioners, since it defines the steps to follow in order to use automatic speaker recognition systems in forensic identification. The methodology considers both the upcoming requirements in forensic science and the common procedures existing in automatic speaker recognition. As a result, three abstraction levels are proposed, namely the discrimination, presentation and forensic levels. Each level has been described in detail, highlighting their main objectives, inputs and outputs. The methodologies for yielding the desired outputs at each level have also been addressed and illustrated with experimental examples, using databases and protocols from widely accepted benchmarks in automatic speaker recognition such as the NIST Speaker Recognition Evaluations (SRE). Moreover, results have been also illustrated by the use of systems and databases used by the Spanish Guardia Civil in real forensic casework. All the techniques presented in this Thesis in order to fulfill the requirements of the proposed hierarchical methodology include some existing approaches found in the literature and also several original contributions.

8.1. Conclusions

Chapter 1 introduced basic concepts about forensic science and criminalistics. The current debate about the scientific foundations of forensic identification is presented, and its consequences are sketched. The basics about evidence evaluation and automatic speaker recognition systems are then introduced. The motivation and the research contributions originated from this Thesis are also described in Chapter 1.

Chapter 2 described related works from the literature concerning automatic speaker recognition and LR–based evidence evaluation using automatic speaker recognition systems. The Dissertation follows with Chapter 3 which describes the databases and experimental protocols used in the Thesis, mainly based on NIST Speaker Recognition Evaluations. The chapter fi-
nally presented the baseline systems used for presenting results in this Thesis, which have been developed at ATVS - Biometric Recognition Group.

The novel hierarchical methodology for forensic automatic speaker recognition is then derived in Chapter 4, considering both the requirements of the coming paradigm shift in forensic science and the common methodologies used in automatic speaker recognition. This leads to the definition of three abstraction levels, namely the discrimination, the presentation and the forensic level. The rest of the Dissertation focuses on the definition and detailed description of each of the proposed abstraction levels in the hierarchy.

Chapter 5 describes the discrimination level. The main objective at this level is yielding a discriminating value, typically a similarity score, in order to distinguish between pairs of speech material coming from the same or from different sources. This level is addressed as sufficient in most automatic speaker recognition applications such as access control or speaker monitoring, according to the literature in the field. First, a definition of discriminating power is given, derived from the automatic speaker recognition literature and benchmarks. Widely accepted performance assessment techniques such as DET plots are then introduced as a measure of discriminating power. The chapter then focuses on the experimental comparison of widespread techniques for improving the discriminating power of the scores in automatic speaker recognition, namely score normalization, session variability compensation and fusion of systems. As a contribution of this Thesis, a novel speaker- and test-dependent technique for score normalization, namely KL-T-Norm, is also proposed at this level. It is experimentally shown that KL-T-Norm outperforms other classical score normalization techniques such as T-Norm, while reducing the computational burden needed for normalizing scores in some scenarios.

In Chapter 6, the presentation level is introduced. At this level, the aim is adapting the scores from the discrimination level in order to compute a likelihood ratio (LR). This value represents the weight of the evidence in the forensic identification process. The use of LR values for forensic reporting was proposed and popularized by the DNA methodology, which has been appointed as the golden standard of scientifically sound procedures in forensic science. First, the chapter introduces the LR-based methodology for evidence evaluation. The meaning and interpretation of the LR value has been described in this Thesis as a degree of support to the fact finder’s prior opinion about the case. After defining its meaning, we address the concept of accuracy of LR values, which is fundamental for the admissibility of forensic techniques in court. According to the statistics literature, the accuracy of a forensic automatic speaker recognition system is defined in terms of two components: first, discriminating power as defined in Chapter 5; and second, calibration of the LR values. We then present performance measures of the accuracy of LR values, from classical representations such as Tippett plots, to recently proposed metrics which consider calibration, such as $C_{\text{llr}}$ and APE plots. It is then empirically demonstrated that an inadequate calibration may lead to highly inaccurate LR values, even if the discriminating power of such LR values is acceptable.

After that, we show how to improve the accuracy of LR values, by means of the experimental comparison of several LR computation approaches. Among them, we propose a novel technique...
for LR computation, namely suspect-adapted LR computation, which demonstrates to improve the discrimination and calibration of the input scores by means of the adaptation to suspect specificities. Moreover, the proposed suspect-adapted technique is shown to be robust to data scarcity. This chapter concludes with an alternative configuration of the proposed hierarchical methodology for forensic automatic speaker recognition in order to consider non-score-based LR computation techniques typical in other forensic areas, and also recently proposed in automatic speaker recognition.

In Chapter 7, the forensic level is presented. Here, the aim is reporting and presenting results in court following the main requirements of the coming paradigm shift in forensic science, identified in Chapter 4. The chapter first recommends several guidelines for forensic reporting. As a conclusion, forensic reporting must consider speech variability, and the conditions of the databases used at every stage in the evidence evaluation process must be clearly stated. Then, the chapter introduces some guidelines about presenting the results from speech evidence in court in a clear and intuitive way, by means of two simulated cases. The first simulated case uses the ATVS-UAM system and NIST SRE 2006 protocol. The second case has been built with systems and databases used by the Spanish Guardia Civil in real forensic casework. These examples illustrate the importance of transparent methods and database collection from field data in order to adequately present results in court. The chapter ends with a generalization of the proposed hierarchical methodology to glass and paint evidence, with experimental results illustrating its adequacy to forensic disciplines other than speech.

8.1.1. Main results and contributions

Summarizing, the main results and contributions obtained from this Thesis are:

- The proposed hierarchical methodology for forensic automatic speaker recognition, with two different configurations: three abstraction levels for the typical score-based automatic speaker recognition architecture; and two abstraction levels for general LR-based evidence analysis techniques, not necessarily based on an intermediate score computation step.

- The identification of the interrelated needs derived from the coming paradigm shift in forensic identification science, namely: transparency, testability, accuracy and common procedures.

- The clear definition of a discrimination level in forensic automatic speaker recognition, sufficient for many applications of score-based automatic speaker recognition systems, and the contribution to the improvement of the discrimination performance by means of novel score normalization methods.

- The definition of the accuracy of forensic automatic speaker recognition as empirical cross-entropy \(ECE\), and the proposed information-theoretical framework for its assessment.
The identification of the need of a presentation level in order to adapt the output scores from an automatic speaker recognition system to a Bayesian probabilistic framework, and the proposed suspect-adapted technique for obtaining accurate LR values.

The review of the reporting considerations in the light of the coming paradigm shift and the particular characteristics of speech.

The demonstration of the adequacy of the proposed methodology to other LR-based forensic disciplines such as glass and paint analysis.

8.2. Future Work

A number of research lines arise from the work carried out in this Thesis. We consider the following as especially interesting:

- Considering the application of the proposed methodology to other biometric traits such as signature, handwriting or face recognition, which have interest among the forensic community [Dessimoz and Champod, 2007]. Several works in that direction have been presented by ATVS – Biometric Recognition Group, such as Gonzalez-Rodriguez et al [2005a]; Ramos-Castro et al [2005], but an extension is needed in order to adapt such biometric traits to the proposed hierarchical methodology.

- Using the proposed hierarchical methodology for other forensic disciplines such as glass, paint or drugs analysis. Although some work has been already reported [Ramos et al, 2007], it is necessary to continue working on the extension of databases of glass and paint chemical profiles, as well as the comparison of different LR computation methods for evidence evaluation. The application to other forensic disciplines such as handwriting examination or drug analysis is also a proposed line of research.

- Combining traditional and automatic speaker recognition approaches. As it is sketched by Gonzalez-Rodriguez et al [2007], many experts agree in the need of combining the traditional (phonetic-acoustic) and automatic approaches for forensic speaker recognition [Künzel, 1994; Rose, 2002]. Their integration into a LR framework and the use of fusion techniques for efficient combination of information are interesting topics to cover in the future.

- Computing LR values directly from the features for automatic speaker recognition, as proposed by Jiang and Deng [2001]; Vogt and Sridharan [2004], and comparing them to other score-based approaches of forensic automatic speaker recognition. Moreover, several techniques for computing LR values from the features have been proposed in other forensic disciplines [Aitken and Taroni, 2004; Aitken et al, 2007; Lindley, 1977], and therefore they can be also applied to forensic automatic speaker recognition.
8.2 Future Work

- Extending the guidelines given in Chapter 7 in order to report and present results in court, with the use of comparative experimental studies using real databases and systems. A preliminary example has been given in this Thesis with systems and databases used by the Spanish Guardia Civil. Mismatching conditions and adequacy of testing protocols should be a key issue of such a future work.

- Integrating recently proposed session variability compensation techniques in the LR framework, and analyzing its impact in forensic realistic speech conditions. Extending such compensation techniques to other forensic disciplines.

- Using measures of speech signal quality in order to aid in the LR computation process. The quality of the speech signal is highly variable in forensic conditions, and it seriously degrades the accuracy of the score and LR computation process. In [Fierrez 2006; Kryszczuk et al. 2007], some measures of quality and reliability of the biometric sample are used for fusing multimodal and multi-algorithm biometric systems. In [Garcia-Romero et al. 2006; Richiardi et al. 2006], the quality of speech signals is used for improving discrimination performance of automatic speaker recognition systems. It is interesting to exploit the quality signal for improving the overall accuracy (discrimination and calibration) in the LR computation process.

- Extending the suspect-adapted approach proposed in this Thesis to other LR computation methods, such as generative kernel density functions, Gaussian mixture modelling or logistic regression. The objective will be exploiting the specificities of the suspect’s control material in order to improve the accuracy of the LR computation technique.

- Applying adaptive methods such as the suspect-adapted LR computation technique proposed here for LR computation using different biometric traits, multimodal biometrics and other forensic disciplines. The objective will be the robustness to data scarcity in background databases.
References


REFERENCES


CAVE. CAVE - the european caller verification project. CAVE project: http://www.ptt-telecom.nl/cave/index.html.


REFERENCES


DNA Advisory Board. Statistical and population genetics issues affecting the evaluation of the frequency of occurrence of DNA profiles calculated from pertinent population database(s). Forensic Science Communications, 2(3), 2000.


Innocence Project. Cardozo School of Law (New York, NY), www.innocenceproject.com. 51 52


LDC. Home page of Linguistic Data Consortium: http://www.ldc.upenn.edu/.


REFERENCES


REFERENCES


SPEECHDAT. Home page of SPEECHDAT project: http://www.phonetik.uni-muenchen.de/SpeechDat.html.


