Extended Abstract for Best Ph.D. Thesis Award: *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*

Daniel Ramos Castro

ATVS - Biometric Recognition Group, Escuela Politécnica Superior
C./ Francisco Tomás y Valiente 11, Universidad Autónoma de Madrid E-28049 Madrid, Spain
daniel.ramos@uam.es

Abstract

This Ph.D. Thesis proposes a complete methodology for the adaptation of automatic speaker recognition technology to forensic evaluation of the evidence. The nature of the problem of forensic automatic speaker recognition is deeply analyzed in the context of the current debate about scientific procedures in forensic disciplines worldwide. Then, a solution for this problem is proposed in the form of a hyerarchical methodology which integrates current standards and state of the art of automatic speaker recognition technology and the requirements and needs of the so-called *coming paradigm shift* in forensic science. The Thesis contributions are supported by numeorus peer-reviwed publications in national and international conferences and jorunals included in ISI-JCR. Also, this Thesis and its constributions have been the recipient of several awards in different national and international contexts. Moreover, the applicability of the Thesis is evidence by the multiple public and private contracts and projects which consider the framework presented here, as well as the impact of the proposed methodologies in important fora such as working groups of the European Network of Forensic Science Institutes.

Index Terms: Forensic speaker recognition, likelihood ratio, calibration, empirical cross-entropy, coming paradigm shift.

1. Nature of the Problem

This Thesis is focused on the use of automatic speaker recognition systems for forensic identification, in what is called forensic automatic speaker recognition [1, 2]. More generally, forensic identification aims at individualization, defined as the certainty of distinguishing an object or person from any other in a given population [3]. This objective is followed by the analysis of the forensic evidence [4], 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 [5, 6]. 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 [6]. 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.

2. Hypotheses to be Tested and Objectives

The Thesis presented, which summarizes the hypothesis to be tested, 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.*

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.

3. Methodology and Relevant Theory

In Chapters 1 and 2, we begin by reviewing related works in the literature concerning automatic speaker recognition and foren-

This Thesis has been funded by a FPI Scholarship from Comunidad Autónoma de Madrid. The Thesis has obtained the Best Student Thesis Award from the Spanish Official College of Telecommunication Engineers (COIT) in 2009. The full Dissertation can be downloaded at http://atvs.ii.uam.es/files/2007_11_28_thesis_daniel_ramos_preprint_v1.pdf

sic 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 [7]. 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 in Chapter 4. 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 [6]. 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 in Chapter 5. The aim at this level is yielding 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 [8], session variability compensation [9] and fusion of systems [10]. A novel score normalization technique, namely KL-T-Norm, is presented as a contribution [11]. We experimentally demonstrate that KL-T-Norm increases the discriminating power of other popular score normalization techniques such as T-Norm [8], as well as it improves its computational efficiency.

Next, the presentation level is introduced in Chapter 6. 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.

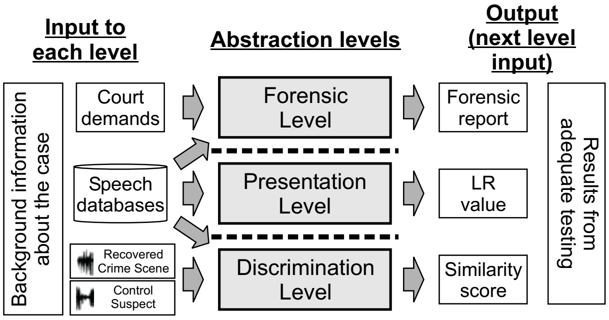


Figure 1: Proposed methodology for forensic evaluation of the evidence using speaker recognition systems, with inputs and outputs of each level in the hierarchy.

Finally, the last level in the hierarchy is described in Chapter 7, 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.

Figure 1 shows the proposed methodology for the use of automatic speaker recognition systems for forensic evidence evaluation. The described hierarchy of levels is shown, as well as the inputs and outputs of each level.

4. Contributions of the Thesis with Indicative References

The Thesis has generated a significant amount of research contributions, evidenced by the number of articles in conferences and journals with ISI-JCR impact factor. As a highlight, the results of the Thesis have received several awards and distinctions, namely:

- Best Ph.D. Thesis Award of the Official College of Telecommunication Engineers (COIT) in 2009.
- IBM Research Best Student Paper Award at the IEEE/ISCA Odyssey 2006 conference, for the article in [12].
- Finalist of the Spanish Network of Speech Technologies (RTTH) Best Journal Article Award, for the work in [11].

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

- **Literature reviews**. Forensic evidence evaluation techniques in automatic speaker recognition [2, 12, 13];. Assessment of forensic speaker recognition systems [13, 14, 12] (IBM Research best student paper award). New requirements in forensic science [13, 2, 12] (IBM Research best student paper award). Automatic speaker recognition [15]. Score normalization for robust speaker verification [16][11] (RTTH best article finalist).

- **Theoretical frameworks.** Theoretical framework for the use of speaker recognition for forensic purposes [2, 12] (IBM Research best student paper award). Theoretical framework for the use of information theory for the assessment of LR values [17, 18, 19, 20].
- **Novel methods.** Novel methods for the use of automatic speaker recognition for forensic identification [2, 21] [12] (IBM Research best student paper award). Novel methods in robust LR computation [21, 14, 22, 23, 24, 20]. Novel methods in the assessment of LR values [17, 18, 19]. Novel methods of score normalization in speaker verification [16][11] (RTTH best article finalist).
- **Improvements in speaker recognition discrimination.** Contribution to the improvement of ATVS-UAM automatic speaker recognition system [2, 25, 14][11] (RTTH best article finalist).
- **New techniques in speaker verification.** New methods for the improvement of automatic speaker recognition discriminating power [26, 16][11] (RTTH best article finalist).
- **New experimental studies.** Experimental studies of automatic speaker recognition systems in the proposed methodology for forensic automatic speaker recognition [13, 2, 19, 12] (IBM Research best student paper award). Robustness in LR-based evaluation of the evidence [21, 14, 27, 22, 23, 24]. Calibration loss effects in forensic speaker recognition [13, 2, 19, 12] (IBM Research best student paper award). Reports on the ATVS-UAM automatic speaker recognition system with forensic applications at NIST SRE and at the NFI/TNO Forensic SRE [13, 2, 25, 26, 14, 12, 27, 16, 28, 22, 23, 24] [11] (RTTH best article finalist). Robust score normalization in speaker verification [16, 28][11] (RTTH best article finalist).
- **Application to other forensic disciplines.** Robust evidence evaluation methods in biometrics [27, 29]. Information theoretical evaluation of LR values coming from glass and paint evidences [17, 18, 19].

5. Results and Analysis

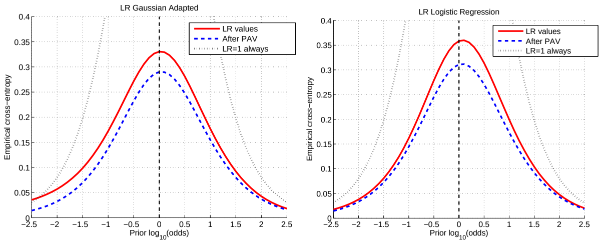
This section presents an analysis of the main results of this Thesis. First of all, the global contribution of the Thesis is the hierarchical methodology for forensic automatic speaker recognition, containing three levels (Figure 1).

The Thesis also clearly defines the discrimination level, where a novel score normalization technique is proposed, namely KL-TNorm. This method improves the discriminating power of systems with respect to the state of the art in test-dependent score normalization, represented by T-Norm [8]. Table 1 illustrates the increase in discriminating power of KL-TNorm in one of the experimental set-ups in the Thesis, which can be seen as a reduction of the Equal Error Rate (EER).

At the defined presentation level, we present and compare several methodologies to compute this evidence weight in terms of likelihood ratios (LR), following the DNA standard. At this level, we define the precision of a set of LR values by means of the Empirical Cross-Entropy (ECE), and its decomposition into discrimination and calibration performance. This methodology allows the presentation of performance results in terms of information theory, and can be illustrated in terms of ECE plots (Figure 2), which constitutes a step forward to make the understanding of the results presented by forensic practitioners easier to the court. Also at this level, we present a novel method for transforming scores from speaker recognition systems into LR values, namely suspect-adapted LR computation, which improves the state of the art of evidence evaluation methods when the amount of speech from the suspect is sparse (Figure 2).

GMM $K = 75$	1c-1c		8c-1c	
	male	female	male	female
EER T-Norm (Av.)	11.14	14.62	7.78	9.57
EER KL-T-Norm	10.76	13.88	7.25	9.12
EER Av. Improvement	3.4%	5.0%	6.8%	4.7%

Table 1: Comparison of EER for TNorm and the proposed KL-Tnorm for the for ATVS GMM system presented in NIST SRE 2005 [7].



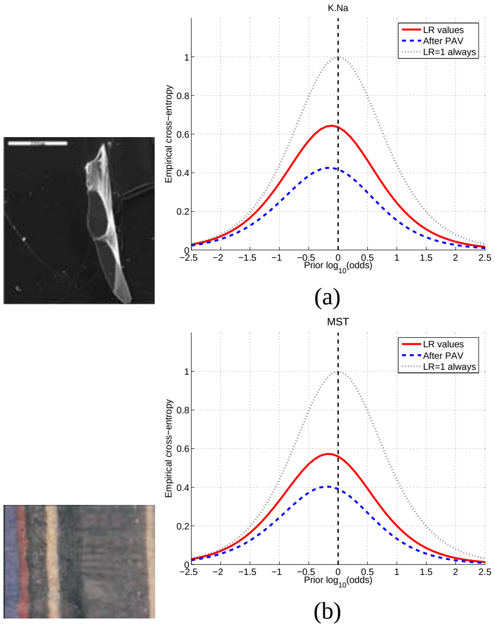
(a) (b)

Figure 2: ECE plots to represent the accuracy of an experimental set of LR values. Red curve denotes information loss, and should be as low as possible. The difference among red and blue curves denote a calibration problem. Comparison of suspect-adapted (a) and logistic regression (b) LR computation.

Finally, we defined a forensic level in the proposed methodology, where the requirements of the *coming paradigm shift* in forensic science are taken into account. The whole methodology is tested by applying it to the procedures followed in the Acoustics Department of the Criminalistics Service of the Spanish Guardia Civil. Moreover, we demonstrate the applicability of the proposed methods to other forensic disciplines, such as Glass or Paint Analysis (Figure 3).

6. Applicability

The main results of this Thesis have been critical as research results and and technology transferred in the context of public and private research projects and contracts. In this sense, we have to highlight the stable collaboration agreement between the ATVS group and the Criminalistics Service of Spanish Guardia Civil. Nowadays, their Acoustics Department are implementing the methodology contributed in this Thesis for forensic evidence evaluation in real casework. Moreover, ATVS, and in particular the author and the advisor of this Ph.D. Thesis, are regular invited members of the Forensic Speech and Audio Analysis Working Group of the European Network of Forensic Science Institutes (ENFSI-FSAAWG), where the experience of Guardia Civil in the deployment of the proposed methods in casework serves as a driving standardization effort for other forensic lab-



(a)

(b)

Figure 3: Application of the proposed methodology based on ECE plots to forensic glass (a) and paint (b) analysis.

oratories and police forces across Europe. Finally, part of the results in this Thesis have been used in technology transfer contracts with Agnitio S. L., and as part of the results of collaboration agreements with the Spanish Ministry of Defense.

7. References

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

[2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

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

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

[5] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*, National Academy of Sciences, 2009.

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

[7] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the Mixer corpora-2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[10] N. Brümmer et al., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[11] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Speaker verification using speaker- and test-dependent fast score normalization," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 90–98, 2007.

[12] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in transparent and testable forensic speaker recognition," in *Proc. of Odyssey*, 2006.

[13] J. Gonzalez-Rodriguez and D. Ramos, "Forensic automatic speaker classification in the coming paradigm shift," in *Speaker Classification*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, 2007.

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

[15] J. Gonzalez-Rodriguez, D. Ramos-Castro, D. Garcia-Romero, and J. Ortega-Garcia, *Reconocimiento del Locutor*, chapter 7, pp. 174–206, RA-MA Editorial, 2005.

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

[17] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, J. Zieba-Palus, and C. G. G. Aitken, "Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation," in *Proceedings of International Workshop on Computational Forensics*, 2007, pp. 411–416.

[18] D. Ramos, "Evaluation of likelihood ratios based on information theory," 2007, Invited presentation at One Day One Topic Seminar and Workshop - Forensic Evidence Evaluation. Institute of Forensic Research, Cracow, Poland. 22nd-23rd June 2007.

[19] D. Ramos and J. Gonzalez-Rodriguez, "Information-theoretical evaluation of likelihood ratios," *Problems of Forensic Sciences*, 2007.

[20] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy analysis of the information in forensic speaker recognition," in *Proc. of Odyssey*, Stellenbosch, South Africa, 2008.

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

[22] J. Gonzalez-Rodriguez, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "On robust estimation of likelihood ratios: The ATVS-UPM system at 2003 NFI/TNO forensic evaluation," in *Proc. of Odyssey*, 2004, pp. 83–90.

[23] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Forensic identification reporting using automatic speaker recognition systems," in *Proc. of ICASSP*, 2003, vol. 2, pp. 93–96.

[24] J. Gonzalez-Rodriguez, D. Ramos-Castro, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Analisis forense de voces dubitadas en la metodologia bayesiana," in *Actas del II congreso de la sociedad espaola de acustica forense (SEAF), Barcelona, Espaa*, 2003.

[25] J. Gonzalez-Rodriguez, D. Ramos-Castro et al., "On the use of high-level information for speaker recognition: the ATVS-UAM system at NIST SRE 2005," *IEEE Aerospace and Electronic Systems Magazine*, pp. 15–21, 2007.

[26] I. Lopez-Moreno, I. Mateos-Garcia, D. Ramos, and J. Gonzalez-Rodriguez, "Support vector regression for speaker verification," in *Proc. of Interspeech*, 2007, pp. 306–309.

[27] D. Ramos-Castro, J. Gonzalez-Rodriguez, C. Champod, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Between-source modelling for likelihood ratio computation in forensic biometric recognition," in *Proc. of AVBPA*, 2005, vol. Springer LNCS-3546, pp. 1080–1089.

[28] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Reconocimiento de locutor multinivel y aplicaciones forenses. el sistema ATVS en NIST SRE 2005," in *Actas del III congreso de la sociedad espaola de acustica forense (SEAF), Santiago de Compostela, Espaa*, 2005.

[29] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems," *Forensic Science International*, vol. 155, no. 2-3, pp. 126–140, 2005.