Suspect-Adapted MAP Estimation of Within-Source Distributions in Generative Likelihood Ratio Estimation

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

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

Abstract

In this paper, a novel suspect-adaptive technique for robust Bayesian forensic speaker recognition via Maximum A Posteriori (MAP) estimation is presented, which addresses Likelihood Ratio (LR) computation in limited suspect speech data conditions obtaining good calibration performance. Robustness is achieved by the use of speaker-independent information, adapting it to the specificities of the suspect involved in the process. Thus, this procedure allows the system to weight the relevance of the suspect specificities depending on the amount of suspect data available via MAP estimation. Experimental results show robustness to suspect data scarcity and stable performance for any amount of suspect material. Also, the proposed technique outperforms other previously proposed non-adaptive approaches. Results are presented as discrimination capabilities (DET plots), distributions of LRs (Tippett plots) and expected cost of wrong decisions over any prior or decision cost (C_{llr}). The use of such evaluation metrics allows us to highlight the importance of LR calibration in the performance of a forensic system.

1. Introduction

In forensic speaker recognition, a court of law may ask for an expert opinion about a questioned recording related to a crime and a given suspect. The aim of a forensic scientist in such a case is to report a *meaningful value* which assesses the *strength of the forensic evidence* in this context of identification of sources [1]. In order to assist forensic experts in criminal trials, the Bayesian framework for evidence evaluation [2, 3] can be applied to forensic speaker recognition by means of automatic systems [4, 5]. In this case, the forensic evidence can be regarded as the human interpretation of all the information that the speaker recognition system can automatically obtain, typically the similarity score between the questioned recording and the suspect speech material. Bayesian interpretation of the forensic evidence using automatic systems has been accomplished both by generative statistical models [4] and discriminative techniques [5]. In this sense, it has been shown in the literature [4] that the the accuracy of LR computation is especially affected by small sample size effects due to suspect data scarcity. Several approaches to achieve robustness against lack of suspect speech have been proposed [4, 6]. In this paper we propose a novel technique which achieves robustness by exploiting speaker-independent information and suspect specificities using an adaptive approach. The proposed technique obtains better discrimination and calibration results compared to other non-adaptive approaches. The paper is organized as follows. Section 2 describes the methodology for LR computation using generative techniques and automatic systems and the related work in the field. The proposed technique, namely within-source distribution estimation via Maximum a Posteriori (MAP) adaptation, is presented in Section 3. In section 4 some results are reported showing the robustness of the technique proposed. Conclusions are drawn in section 5.

This work has been supported by the Spanish Ministry for Science and Technology under project TIC2003-09068-C02-01. The author D. R.-C. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research.

2. Generative Likelihood Ratio Computation Using Automatic Systems

2.1. Bayesian interpretation of forensic evidences

Likelihood Ratios (LR) can be estimated from similarity scores computed by an automatic system [4]. In order to obtain such a value, a probabilistic model based on the odds form of Bayes' theorem and Likelihood Ratio (LR) computation has been shown to be an adequate tool for assisting experts in forensic sciences to interpret evidence [2, 3]. This Bayesian framework for interpretation of the evidence presents many advantages in the forensic context. First, it allows the forensic scientists to estimate and report a meaningful LR value to the court [1], where the numerical LR value means a support to one of the hypotheses involved (e. g., $LR = 6$ means that there is a 6 versus 1 support to one hypothesis respect its opposite). Therefore, this value allows not only to discriminate between suspects, but also to infer posterior probabilities, or *confidences* [5], in order to take decisions in a transparent and scientific way. Second, the role of the scientist is clearly defined, leaving to the court the task of using prior judgements or costs in the decision process [7]. Third, probabilities can be interpreted as degrees of belief [8], allowing the incorporation of subjective opinions in the inference process. Finally, there is an extensive work in the literature related to the evaluation of posterior opinions and LRs as a degree of support of any of the hypotheses involved in the Bayesian inferential process [9]. Moreover, useful evaluation measures of the LR with attractive information-theoretical interpretations can also be found in [9]. There, the LR is evaluated through the C_{llr} metric in an application-independent way, i. e., independently of the different prior opinions and costs involved in the decision process [7].

This evaluation metric has been recently proposed by NIST in next 2006 Speaker Recognition Evaluation (SRE) [10] for the evaluation of speaker recognition systems providing LR values instead of scores. Moreover, C_{llr} can also be interpreted as information delivered from the forensic system to the user in the context of Information Theory.

2.2. Generative Likelihood Ratio Computation

Following this approach for LR computation, we assume that the evidence E is the similarity score between the questioned speech and the suspect material computed with the speaker recognition system at hand. Therefore:

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

where H_p (a given suspect is the author of the questioned recording involved in the crime) and H_d (another individual is the author of the questioned recording involved in the crime) are the relevant hypotheses and I is the background information available in the case. The likelihoods $f\left(x|\, H_p, I\right)$ and $f\left(x|\, H_d, I\right)$ are respectively known as the within- and between-source probability density functions (pdf). Between-source pdf is modelled from scores assuming that H_d is true. These *non-target* or *impostor* scores are obtained comparing the questioned speech under analysis with a population of individuals. On the other hand, within-source distribution is estimated from scores assuming that H_p is true. These within-source scores are obtained comparing different utterances from the suspect speech material, and therefore they will be considered as *target* or *genuine* scores. See [4] for details.

One of the main problems in within-source estimation is related with the suspect speech data scarcity [4]. In [6], a framework is proposed assuming that an accurate model of the within-source distribution for a given suspect can be obtained using target scores from different individuals in the same conditions. However, it has been shown that, even in the same conditions, the target scores coming from different speakers may present different distributions [11]. Therefore, accuracy in within-source estimation may be improved by exploiting suspect-specific scores, because the H_p condition claims that the suspect *and no other individual* is the author of the questioned recording. In [4] a different approach is proposed, namely Within-source Degradation Prediction (WDP). This technique combines suspect target scores with between-source distribution information to predict score variability not present in the suspect data. Experiments presented in [4] show excellent performance when limited suspect data is available. However, this optimization technique, despite improving the discrimination performance of the system, introduces errors in the posterior probabilities inferred from the LR. This is because WDP aims at fixing the within-source distribution without considering the actual (and unknown) suspect data it claims to represent. Therefore, the predicted within-source pdf will not represent the actual distributions, and thus the technique will incur a calibration loss. Then the information provided by the system is suboptimal. This effect is solved by the technique proposed below, and is discussed in depth in another recent work from the authors [12].

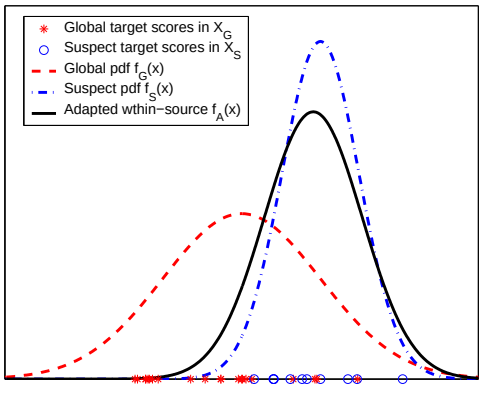

Figure 1: MAP adaptation example with $r = 1$. Global distribution (dashed) representing speaker-independent information is adapted to suspect data.

3. MAP Adaptation for Suspect-Adapted Within-source estimation

In this work, a novel adaptive approach to within-source computation is proposed, which exploits both general speaker-independent variability and suspect specificities. Our strategy is based on the adaptation of the speaker-independent target score distribution to the suspect target scores via MAP estimation [13]. Let $X_G = \{x_{G1}, \ldots, x_{GN}\}$ be a set of *global* target scores computed using speech from speakers other than the suspect. Let $X_S = \{x_{S1}, \ldots, x_{SM}\}$ be a set of *suspect* target scores obtained from the suspect speech involved in the trial. First, using Maximum Likelihood and assuming Gaussian distributions, we estimate the pdfs $f_G\left(x\right) = N\left(\mu_G, \sigma_G\right)$ and $f_S\left(x\right) = N\left(\mu_S, \sigma_S\right)$ from X_G and X_S respectively. $f_G\left(x\right)$ represents the variability of target scores between speakers. On the other hand, $f_S\left(x\right)$ represents the suspect target scores variability. Assuming $f\left(x|\, H_p, I\right) \equiv f_A\left(x\right) = N\left(\mu_A, \sigma_A\right)$ (see Equation 1), we compute the within-source distribution parameters using MAP adaptation as follows [13]:

$$\begin{aligned} \mu_A &= \alpha\, \mu_S + \left(1 - \alpha\right) \mu_G \\ \sigma_A^2 &= \alpha\, \left(\sigma_S^2 + \mu_S^2\right) + \left(1 - \alpha\right) \left(\sigma_G^2 + \mu_G^2\right) - \mu_A^2 \end{aligned} \tag{2}$$

The *adaptation coefficient* α is defined as

$$\alpha = \frac{M}{M + r} \tag{3}$$

and depends on: *i)* the number of suspect scores M and *ii)* a fixed relevance factor r. It is observed that when M is small, the algorithm gives more importance to global data X_G. As more suspect scores are available, the adapted within-source distribution will be more adjusted to the suspect data X_S. Note that if $r = 0$ then $f_A\left(x\right) = f_S\left(x\right)$. On the other hand, if $r \to \infty$ then $f_A\left(x\right) \to f_G\left(x\right)$ and the resulting within-source will be speaker-independent as in [6]. Figure 1 illustrates this technique for $r = 1$ from 10 suspect scores and 20 global target scores.

4. Experiments

Experiments have been performed using the evaluation protocol proposed in NIST 2005 SRE [14]. The database used in this evaluation has been extracted from the MIXER corpus [15],

and includes different communication channels, handsets, microphones and languages. The evaluation protocol defines different training and testing conditions. We carry out our experiments using the ATVS GMM-MAP-UBM system submitted to NIST 2005 SRE. KL-Tnorm technique, an efficient and adaptive speaker- and test-dependent score normalization technique, has been used to normalize scores [16]. Results are presented for the 8 conversation side training and 1 conversation side testing task (8c-1c). Each conversation side has an average duration of 2.5 minutes of speech after silence removal. In this condition, more than 250 speakers are involved, and more than 23000 trials are performed. Details can be found in the NIST 2005 SRE Evaluation Plan in [14]. In order to obtain each suspect's target score set X_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 NIST 2004 SRE database was selected. Trials performed using this development set follow the NIST 2004 SRE protocol. The global target score set X_G consists of all the target scores in this development set. As the database used in NIST 2004 was also a different subcorpus of MIXER, X_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. KL-Tnorm is performed in the following experiments through adaptive selection of speaker-dependent 75-models cohorts.

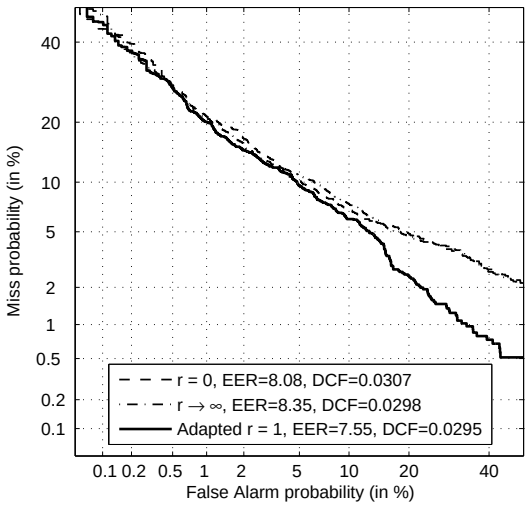

Figure 2: DET plots of LR values for $M = 2$ and speaker-dependent ($r = 0$), speaker-independent ($r \to \infty$) and speaker adapted ($r = 1$) with NIST 2005 SRE data in the selected subset of 8c-1c condition.

4.1. Performance Evaluation

The performance of the system will be presented in three different ways: *i)* DET plots [17] are used to measure the discrimination capabilities of the computed LRs; *ii)* Tippett plots [4], as used in NFI/TNO 2003 Forensic SRE [18], present the distributions of the LRs of each system for target and non-target trials, showing the actual values of the LR values and the rates of misleading evidence for each hypothesis; and *iii)* C_{llr} [9] gives the expected cost of taking wrong decisions when using the forensic system, averaged over a wide range of applications, i.e., prior judgements and costs. Because of its interesting properties, C_{llr} has been selected in NIST 2006 SRE as an evaluation metric for systems delivering LRs instead of scores [10].

DET curves measure the discrimination performance of the speaker recognition system in all operating points. However, the performance of the LRs not only depends on their capability of discriminating among speakers, but in the actual values of the posterior probabilities (or *confidences* [5]) which are inferred, as meaningful values are required (see Section 1). Tippett plots illustrate these values as cumulative distributions of LRs under H_p and H_d plotted together. Here, the rate of misleading evidence, i. e., the proportion of $LRs > 1$ under H_d and $LR < 1$ under H_p, is highlighted as a performance measure, but a single scalar value would be desirable in order to rank and compare overall accuracy of systems. In this sense, C_{llr} *is* a scalar value which is defined as:

$$C_{llr} = \frac{1}{N_{H_p}} \sum_{i\,for\,H_p=true} log_2\left(1+\frac{1}{LR_i}\right) + \frac{1}{N_{H_d}} \sum_{j\,for\,H_d=true} log_2(1+LR_j) \qquad (4)$$

where N_{H_p} and N_{H_d} are respectively the number of LRs in the evaluation set for H_p or H_d true. As it can be seen in Equation 4, a hypothesis-dependent logarithmic cost function is applied to LRs in the test set. Therefore, C_{llr} penalizes $LR > 1$ values when H_d is true and $LR < 1$ values when H_p is true. Also, high LR values when H_p is true and low LR values when H_d is true will provide a lower C_{llr}, and therefore a better performance.

C_{llr} can also be interpreted in an information-theoretical way. Given a system which outputs LRs, $1-C_{llr}$ measures the amount of actual information that is delivered from the system to the user (in our case, the fact finder) assuming a maximum entropy prior (in our binary case, $P(H_p) = P(H_d) = 1/2$). So, the lower the C_{llr} value, the higher the information delivered from the system to the fact finder. Moreover, C_{llr} includes two different measures: *i)* the loss in accuracy because of the discrimination capabilities of the system (i. e., a bad *refinement* [19]) and *ii)* a penalty to LR values which would lead to unreliable or misleading confidences (which is known as a lack of *calibration* [19, 9]). The reader may consult [9] for a detailed description of the effects of calibration in automatic speaker recognition systems and [12] for its application to forensic speaker recognition.

4.2. Results

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 maintaining the rest of conditions. Figure 2 shows the performance of the proposed technique in terms of DET plots in suspect data scarcity conditions ($M = 2$). It is observed that the speaker-adapted within-source estimation technique outperforms discrimination capabilities of speaker-dependent ($r = 0$) and speaker-independent ($r \to \infty$) methods for all operating points. However, this improvement is not so significant for low False Alarm rates (and DCF as defined by NIST [14]). But the performance of the LRs depends not only in their discrimination performance, but also in their actual values. This fact is illustrated by comparing Figure 2 and Figure 3, where the performance of the suspect-adapted within-source estimation technique in terms of Tippett plots is shown. We observe that suspect-adapted within-source computation ($r = 1$) presents lower rates of misleading evidence when H_p is true, having similar rate of misleading evidence when H_d is true.

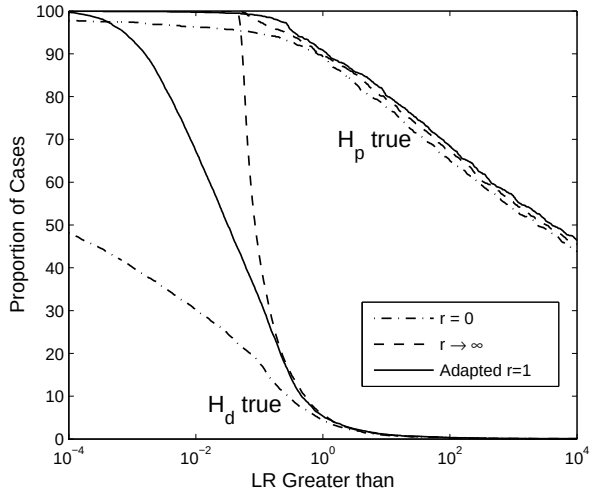

Figure 3: Tippett plots for $M = 2$ and speaker-dependent ($r = 0$), speaker-independent ($r \to \infty$) and speaker adapted ($r = 1$) in the same experiment of Figure 2.

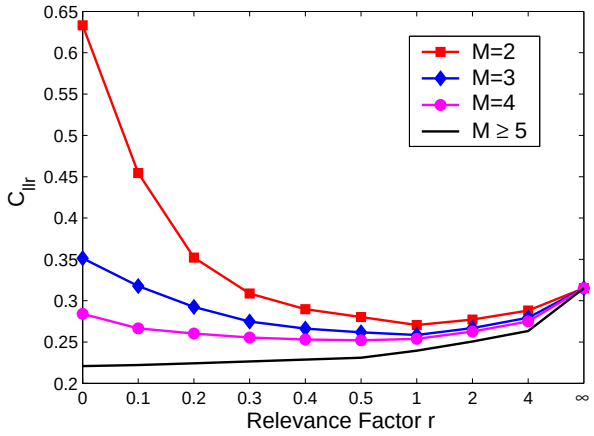

Figure 4: C_{llr} for suspect-adapted LR computation with different amount of suspect scores M and different relevance factors r for the selected subset of the 8c-1c condition in NIST 2005 SRE.

On the other hand, speaker-dependent within-source estimation ($r = 0$) when there is a lack of suspect data leads to seriously misleading LR values under H_p (Figure 3). This lack of calibration [19, 9], which is not observable in a DET plot, represents a critical issue in forensic speaker recognition systems. This important idea is out of the scope of this paper and is deeply addressed in [12].

Performance of the system for different relevance factors r (Equation 3) and different number of suspect target scores M is shown in Figure 4. 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 estimation. 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 $r = 2$.

In order to complete the analysis, Tippett plots in Figure 5 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 4: as M decreases, the performance of the system is similar, especially in the sense of misleading LR values, i. e., $LR > 1$ values when H_d is true and $LR < 1$ values under H_p.

5. Conclusions

This paper has presented a novel, robust adaptive generative Likelihood Ratio (LR) computation technique for addressing Bayesian forensic speaker recognition using automatic systems. The proposed method adapts within-source distribution from a speaker-independent distribution to the suspect target scores using MAP estimation. The presented technique has been shown to be robust against data scarcity and to achieve stable performance when the amount of suspect data grows, while outperforming both speaker-dependent and speaker-independent

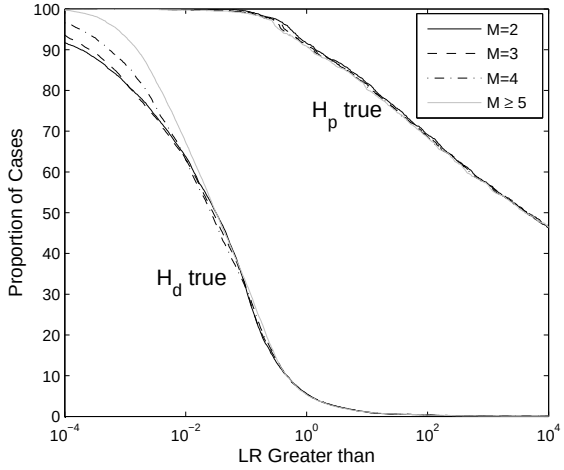

Figure 5: Tippett Plots for suspect-adapted LR computation with different amount of suspect scores M and $r = 1$ for the $r = 1$ case in Figure 4.

non-adaptive approaches, both in terms of DET plots as a measure of discrimination performance, in terms of Tippett plots as a representation of the actual LR values and in terms of application-independent expected cost of wrong decisions (C_{llr}). Finally, the need for calibration in the performance analysis of forensic speaker recognition systems has also been pointed out.

6. References

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

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

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

[4] 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.

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

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

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

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

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

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

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

[12] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood Ratio Calibration in a Transparent and Testable Forensic Speaker Recognition Framework," accepted in *Odyssey 2006*.

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

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

[15] J. P. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. F. Martin, and M. A. Przybocki, "The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation," in *Proc. of Odyssey 2004*, pp. 29–32.

[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] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of decision task performance," in *Proc. of Eurospeech 97*, pp. 1895–1898.

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

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