D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy Analysis of the Information in Forensic Speaker Recognition", to appear in Proceedings of IEEE Odyssey, January 2008.

Cross-entropy Analysis of the Information in Forensic Speaker Recognition

Daniel Ramos and Joaquin Gonzalez-Rodriguez

Biometric Recognition Group - ATVS, EPS, C./ Francisco Tomas y Valiente 11,
Universidad Autonoma de Madrid E-28049 Madrid, Spain
{daniel.ramos, joaquin.gonzalez}@uam.es

Abstract

In this work we analyze the average information supplied by a forensic speaker recognition system in an information-theoretical way. The objective is the transparent reporting of the performance of the system in terms of information, according to the needs of transparency and testability in forensic science. This analysis allows the derivation of a proper measure of goodness for forensic speaker recognition, the empirical cross-entropy (ECE), according to previous work in the literature. We also propose an intuitive representation, namely the ECE plot, which allows forensic scientists to explain the average information given by the evidence analysis process in a clear and intuitive way. Such representation allows the forensic scientist to assess the evidence evaluation process with independence of the prior information, which is province of the court. Then, fact finders may check the average information given by the evidence analysis with the incorporation of prior information. An experimental example following NIST SRE 2006 protocol is presented in order to highlight the adequacy of the proposed framework in the forensic inferential process. An example of the presentation of the average information supplied by the forensic analysis of the speech evidence in court is also provided, simulating a real case.

1. Introduction

Information theory was proposed in the middle of the 20th century as a standard for measuring and presenting information [1]. After more than 50 years, the applications of information theory have been remarkable in many fields like physics, probability theory and economics [2]. Under this framework, the uncertainty about an unknown variable is quantified by a magnitude called *entropy*. Additional knowledge about other known variables under study will contribute to the reduction of the entropy, and therefore, the information about the unknown variable will be increased.

Recently, information theory has been proposed in order to assess the goodness of automatic speaker detection [3, 4]. Such techniques assume that the system yields likelihood ratios (LR) as a degree of support to any of the hypotheses involved in the detection process. Although such 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 two speech material being compared come from the same speaker or not.

In forensic speaker recognition, the LR approach has been proposed for reporting the weight of the evidence in court [5, 6, 7]. Moreover, the uprising requirements in forensic science require the use of scientifically sound procedures for clearly stating the accuracy of the techniques in use. For instance, in a given case and according to Daubert rules or similar criteria [8], the fact finder may demand a test in order to clarify the accuracy of the LR computation technique used for evidence analysis, whose results are to be presented in court. Such assessment will contribute to the decision of the fact finder about the admissibility of the evidence analysis process. In order to fulfill this requirements, in this paper we propose the use of information-theoretical magnitudes for assessing the accuracy of the LR values computed by forensic speaker recognition systems. We will consider that the evidence analysis gives information to the fact finder about the value of the hypothesis involved in the case. The proposed assessment framework measures how good the forensic system is extracting such information, and allows the forensic scientist to present it in court in a clear and transparent way. The importance of transparent reporting of the performance of forensic techniques has been also recently highlighted for forensic speaker recognition [7].

The aim of this paper is identifying and characterizing the information supplied by the weight of the speech evidence computed by the forensic system, considering the requirements of the so-dubbed *coming paradigm shift* in forensic identification [8]. The reduction of the uncertainty gives a measure of the expected information that the evaluation of the evidence delivers to the decision process in a forensic case, and it is modelled in terms of entropy and divergence [2]. These magnitudes will be integrated in a LR-based framework adopted from forensic DNA analysis [7]. In particular, a clear distinction is made between the information sources given by the analysis of the evidence, province of the forensic scientist, and the rest of information in the case, province of the court. A novel performance representation is proposed, namely the ECE *plot*, which integrates previous approaches and gives a clear an 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.

The paper is organized as follows. Section 2 introduces the problem of the assessment of decisions in forensic science and reviews some approaches found in the literature. In section 3, the proposed measure of accuracy, namely *empirical cross-entropy* (ECE), is derived, as well as its interpretation. In Section 4 the ECE plot is presented as a useful performance representation suited for forensic speaker recognition, discussing its relationship to other performance measures already proposed. In Section 5 an experimental example is reported, which illustrates the adequacy of the proposed methods for forensic cases. The section is completed with a simulation of a real case, where the ECE value is reported as a measure of performance according to the requirements of Daubert and similar criteria. Finally, conclusions are drawn in Section 6.

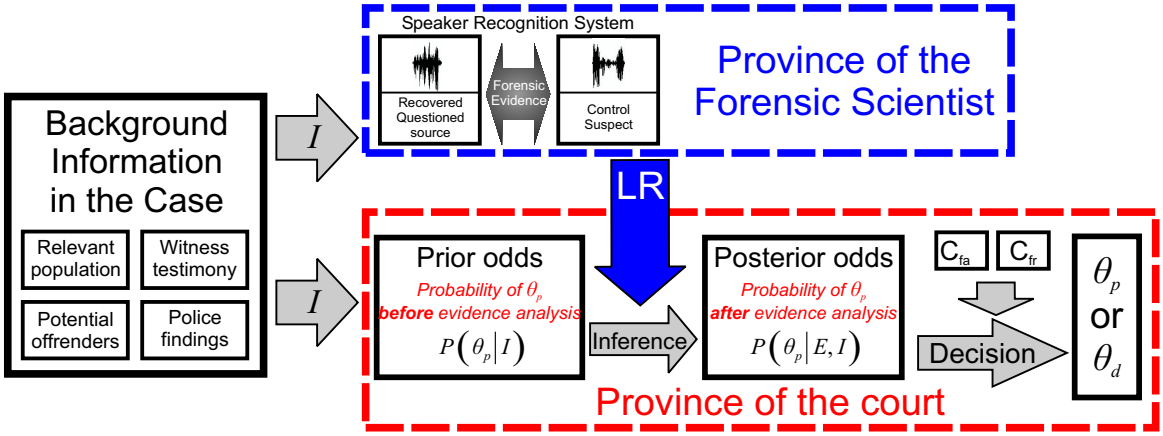


Figure 1: Elements in the decision process using LR-based forensic speaker recognition.

2. Cost-based evaluation of forensic speaker recognition systems

The LR framework for evidence analysis is summed up here [9, 10]. Consider the forensic speech evidence as the comparison of a *recovered* speech sample (of unknown source) and a *control* sample (usually from a suspect). Such comparison will be referred to as a *trial*. Bayes' theorem then allows the following inference:

$$\frac{P(\theta_p \mid E, I)}{P(\theta_d \mid E, I)} = LR \cdot \frac{P(\theta_p \mid I)}{P(\theta_d \mid I)} \quad (1)$$

where θ_p (the control and the recovered samples come from the same source) and θ_d (the control and the recovered samples come from different sources) are typically the relevant hypotheses and I is the background information available in the case. The likelihood ratio (LR) is defined as:

$$LR = \frac{p(E|\theta_p, I)}{p(E|\theta_d, I)} \quad (2)$$

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 distinguish two magnitudes: i) 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 in the case I; and ii) the LR (Equation 2), computed by the forensic scientist [4, 11, 7]. The background information I may include not only circumstantial information in the case (such as witness testimony or police investigations), but also the analysis of other forensic evidences (such as glass fragments, paint flakes, etc.). Such two magnitudes allow the fact finder to infer a posterior probability for each hypothesis $P(\theta_p \mid E, I) = 1 - P(\theta_d \mid E, I)$, which considers both I and the evidence evaluation from the forensic scientist. The background information about the case I will be eliminated from the notation for simplicity from here thereafter, but it will be assumed that all the probabilities are conditioned to I. Thus, we will express prior and posterior probabilities of θ_p respectively as $P(\theta_p)$ and $P(\theta_p|E)$, and similarly for θ_d.

In order to take a decision according to Bayesian theory [12], the fact finder would have to use the posterior and also some decision costs. These costs represent penalties for each type of error in each binary decision, namely false acceptance cost (C_{fa}) and false rejections (C_{fr}). The elements in this inferential process are shown in Figure 1. Ideally, computing the LR value would allow the fact finder to take Bayes decisions, which are known to be optimal in a cost sense [12]. However, unavoidable and realistic imperfections in the computation of the LR values will degrade the optimality of the decisions taken by the fact finder.

2.1. Cost-based evaluation

In order to evaluate the goodness of the fact finder's decisions, 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 θ_p is true, and a set of non-target scores, for which θ_d is true. The results of such a forensic test can then be evaluated in a cost-based way, as the one proposed by the American National Institute of Standards and Technology (NIST) in their Speaker Recognition Evaluations (SRE) since 1996 [13]. Thus, the mean cost is defined as:

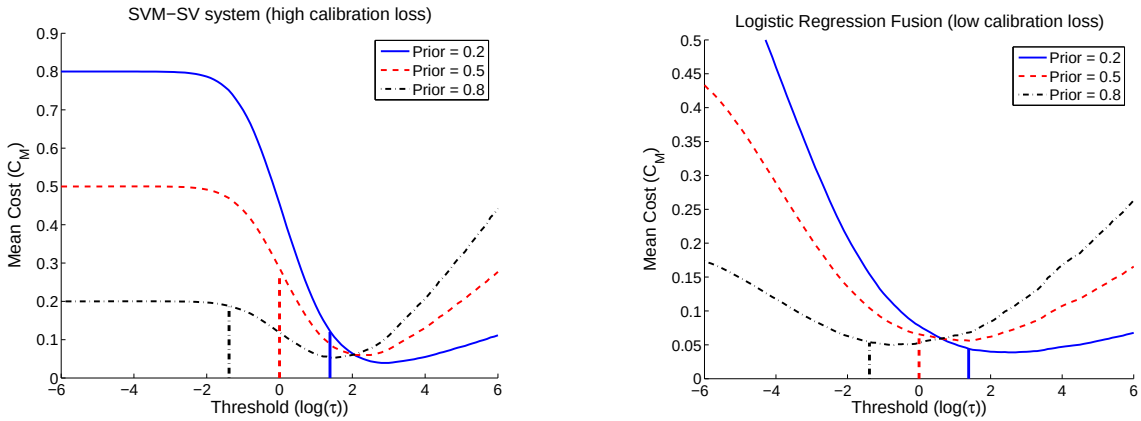
$$C_M = P_{fr}(\tau) \cdot C_{fr} \cdot P(\theta_p) + P_{fa}(\tau) \cdot C_{fa} \cdot P(\theta_d) \quad (3)$$

where $P_{fr}(\tau)$ and $P_{fa}(\tau)$ are the false rejection and false acceptance probabilities of the speaker recognition system, dependent on the decision threshold τ. Assuming that the system yields LR values, τ is defined as:

$$\begin{array}{ll} LR > \tau & : \text{Decide } \theta_\text{p} \\ LR < \tau & : \text{Decide } \theta_\text{d} \end{array} \quad (4)$$

Also, C_{fr} and C_{fa} are the costs respectively applied to each false rejection or false acceptance. $P(\theta_p)$ and $P(\theta_d)$ are the prior probabilities defined in Equation 1. In a forensic context both costs and priors are independent of the forensic system, and the fact finder should state their values, according to the circumstances of each case (I) [10, 9]. For instance, in a case where the fact finder thinks that, in the light of I, there is a 10% of probability that the suspect is the author of the questioned, then it should happen that $P(\theta_p) = 0.1$.

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, it is possible to find a value of the threshold (not necessarily unique), namely τ^*, which leads to a minimum value of the mean cost. We will say



(a) (b)

Figure 2: Value of C_M (Equation 3) for different decision thresholds. (a) SVM-SuperVector system (high calibration loss) and (b) Logistic regression fused system (low calibration loss). $C_{fr} = C_{fa} = 1$. Bayes thresholds (Equation 5) are shown as vertical lines.

that a system is calibrated [3, 14] for given prior and cost values if the decision threshold determines a pair of $P_{fr}(\tau^*)$ and $P_{fa}(\tau^*)$ probabilities which minimize C_M, i.e., when $\tau = \tau^*$. The difference between the optimum value of C_M at τ^* and the value of C_M determined by the selected threshold τ is known as *calibration loss* [3]. As an example, in Figure 2(a) the value of C_M for a range of thresholds τ 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 for the mean cost, which is strongly dependent on the prior probabilities.

However, for a given forensic case the priors and costs are province of the court and may not be even known by the forensic scientist. Moreover, each forensic case is unique, and in general the priors and costs may vary among forensic cases. Hence, if the priors change, the optimum threshold τ^* for the original priors and costs will not be optimum anymore for the new priors and costs, as it is observed in Figure 2(a). Thus, C_M may dramatically increase because this lack of calibration, which is dependent on the value of the prior and the costs. Fortunately, according to Bayes decision theory [12], if the speaker recognition system computes LR values (Equation 2) then the optimum threshold for decision making, commonly known as the Bayes threshold, is given by:

$$\tau_B = \frac{C_{fa} \cdot P(\theta_d)}{C_{fr} \cdot P(\theta_p)} \tag{5}$$

Thus, in order to obtain an optimal value of the mean cost for any prior or cost two conditions are necessary: i) computing a LR value from the score; and ii) setting the Bayes threshold according to the priors and the costs (Equation 5). The former condition should be accomplished by the forensic system, whereas the latter condition should be the duty of the fact finder.

Figure 2 illustrates the effects of calibration. Two different systems are shown, and in both cases the value of C_M is represented for a range of thresholds τ with the constraint that $C_{fr} = C_{fa} = 1$. Figure 2(a) shows a system where decisions are taken directly from the scores , i.e., LR values have not been computed from the scores. In this case, it can be said that the score is used as a LR value [1]. Bayes thresholds (τ_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, selecting the Bayes threshold leads to a suboptimal value of C_M for all cases. However, Figure 2(b) shows a system where LR values have been computed from the scores, and it is shown that τ_B is near from the optimum for all the presented values of $P(\theta_p)$.

Figure 2(b) also shows that the optimality of τ_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 τ_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 τ_B may not be optimal anymore, and therefore a calibration loss will occur. Moreover, in forensic applications, where the value of the priors and the decision costs may be different case to case, it is mandatory to measure the goodness of the computed LR values for any value of the prior and the decision costs.

2.2. Application-independent evaluation

A solution to this problem has been proposed in [3] for speaker recognition, and since 2006 adopted by NIST in their Speaker Recognition Evaluations (SRE) [13]. The values of the priors and the costs in [3] determine an *application*. The measure of accuracy proposed there, namely C_{llr}, is independent of the application, being computed as:

$$\begin{aligned} C_{llr} &= \frac{1}{2 \cdot N_p} \sum_{i \in \text{targets}} log_2\left(1 + \frac{1}{LR_i}\right) \\ &+ \frac{1}{2 \cdot N_d} \sum_{j \in \text{nonTargets}} log_2(1 + LR_j) \end{aligned} \tag{6}$$

where N_p and N_d are respectively the number of target and non-target scores in the evaluation set. Thus, two averages are performed over two different logarithmic function of the scores: one for targets and one for non-targets. In [3] it is demonstrated that C_{llr} is the mean of C_M over all possible values of the decision costs, fixing $P(\theta_p) = 0.5$. Thus, it is expected that optimizing C_{llr} will improve the calibration of the scores for any possible value of the decision costs at $P(\theta_p) = 0.5$ [3].

[1] Scores and LR values are always considered to lay in a common domain. Therefore, if the scores lay in the $(-\infty, \infty)$ range, as it is usual in speaker recognition, then they will be considered $\log(LR)$ values in order to use them with Equation 1.

3. Information-theoretical evaluation

In this section we will derive an information-theoretical generalization of C_{llr}, namely *Empirical Cross-Entropy* (ECE) which measures the accuracy of the LR values in terms of average information loss. ECE is in essence a normalized version of other measures proposed in the literature for application-independent evaluation of speaker detection, such as $U_{\log}$ [15]. Moreover, another normalized version of ECE, namely *normalized cross-entropy* (NCE), has been already proposed in the literature for forensic speaker recognition [4] and in NIST Speech Recognition and Rich Transcription evaluations [16].

3.1. Uncertainty and information

Information theory [1, 2] states that the information obtained in an inferential process is determined by the reduction of the entropy, which measures the uncertainty about a given unknown 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 variable $\theta = \{\theta_p, \theta_d\}$.

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 (I) as defined in Section 2. With this available knowledge, the entropy of the hypothesis, namely *prior entropy* or *entropy of the prior* is determined by the following expression [2]:

$$H_P(\theta) = -\sum_{i\in\{p,d\}} P(\theta_i)\log_2 P(\theta_i) \tag{7}$$

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 θ.

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. In a given forensic case, such LR value may or may not reduce the uncertainty about the hypothesis variable. However, it can be demonstrated [2] that the expected value of the entropy of the posterior probability over all possible values of the evidence E cannot be greater than the prior entropy. This expected value is the *posterior entropy*, computed as [2]:

$$H_P(\theta|E) = -\sum_{i\in\{p,d\}} P(\theta_i)\int_{-\infty}^{\infty} p(e|\theta_i)\log_2 P(\theta_i|e)\,de. \tag{8}$$

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

The expected information supplied by the evidence analysis is illustrated in Figure 3. There, it is represented that knowledge about the evidence will never increase the expected uncertainty about the hypotheses over all possible values of the evidence [2]. However, the computation of Equation 8 is usually non-practical, as it requires the knowledge about the likelihoods $p(e|\theta_i)$ computed by the system. Such likelihoods may not be known in general, e.g., if discriminative LR computation techniques are used as in [4, 3, 7]. Moreover, even when the $p(e|\theta_i)$ 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 (e.g. mismatch between training and evaluation conditions).

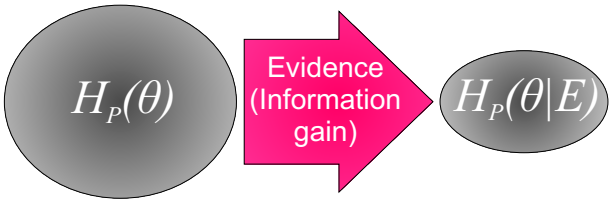


Figure 3: Expected reduction of uncertainty (information gain) due to evidence analysis, over all possible values of the evidence. The area of the ellipses represent entropy, i.e., uncertainty.

A solution to this problem has been proposed in the literature [15, 3, 16] by comparing the posterior probabilities computed using the forensic system with a *reference* probability distribution. The letter P (p for pdfs) will denote probabilities obtained using the forensic system and the letter Q (q for pdfs) will denote reference probabilities. This eliminates the dependence of the posterior and the likelihood inside the integral in Equation 8, leading to the cross-entropy:

$$H_{Q\|P}(\theta|E) = -\sum_{i\in\{p,d\}} Q(\theta_i)\int_{-\infty}^{\infty} q(e|\theta_i)\log_2 P(\theta_i|e)\,de. \tag{9}$$

It can be demonstrated that the cross-entropy (Equation 9) may be decomposed into:

$$H_{Q\|P}(\theta|E) = H_Q(\theta|E) + D_{Q\|P}(\theta|E) \tag{10}$$

where $D_{Q\|P}(H|E)$ is the well-known Kullback-Leibler (KL) divergence between the system's posterior distribution and the reference distribution [2] for all possible values of the evidence, defined as:

$$D_{Q\|P}(\theta|E) = \sum_{i\in\{p,d\}} Q(\theta_i)\int_{-\infty}^{\infty} q(e|\theta_i)\log_2 \frac{Q(\theta_i|e)}{P(\theta_i|e)}de. \tag{11}$$

Thus, the 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.
- $D_{Q\|P}(\theta|E)$, 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 (Equation 9).

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

The computation of the cross-entropy using Equation 9 may be tedious if possible. However, an empirical approximation is used here. Given a target and a non-target set of LR values from forensic testing, we can obtain two target and non-target

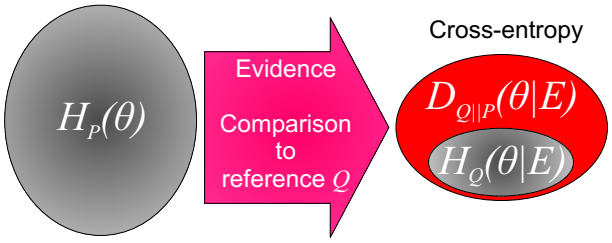


Figure 4: The cross-entropy consists of the posterior entropy of the reference (inner ellipse, uncertainty) plus the divergence between the reference and the posterior probability of the system (red darker area in the outer ellipse, information loss).

sets of posterior probabilities using Equation 1, assuming that the prior probabilities are known. Therefore, we can average the expectations in Equation 9, supposing the law of the large numbers holds, obtaining:

$$\begin{aligned} ECE = & - \frac{Q(\theta_p)}{N_p} \sum_{i \in \text{targets}} \log_2 P(\theta_p | e_j) \\ & - \frac{Q(\theta_d)}{N_d} \sum_{j \in \text{nonTargets}} \log_2 P(\theta_d | e_j) \end{aligned} \quad (12)$$

where:

$$H_{Q||P}(\theta|E) \simeq ECE \quad (13)$$

This value will be our evaluation objective, namely *empirical cross-entropy* (ECE), which is equivalent to the already proposed NCE [16, 4] and $U_{\log}$ [15]. The posterior probability is dependent on the prior probability $P(\theta_p)$ and the LR value, since:

$$P(\theta_p|E) = \frac{LR \cdot \frac{P(\theta_p)}{P(\theta_d)}}{1 + LR \cdot \frac{P(\theta_p)}{P(\theta_d)}}. \quad (14)$$

Then, ECE can be expressed as:

$$\begin{aligned} ECE = & \frac{Q(\theta_p)}{N_p} \sum_{i \in \text{targets}} \log_2 \left(1 + \frac{1}{LR_i \cdot \frac{P(\theta_p)}{P(\theta_d)}}\right) \\ & + \frac{Q(\theta_d)}{N_d} \sum_{j \in \text{nonTargets}} \log_2 \left(1 + LR_j \cdot \frac{P(\theta_p)}{P(\theta_d)}\right) \end{aligned} \quad (15)$$

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, the forensic scientist is allowed to compute and represent ECE for a range of prior probabilities, without assuming a particular value for $P(\theta_p)$. Then, the fact finder can compute the ECE value for the particular prior in a given case.

Figure 4 illustrates the information loss measured by cross-entropy in terms of its decomposition (Equation 10). As the prior is taken as a parameter, then $H_P(\theta) = H_Q(\theta)$. Therefore, from Equation 15, it is straightforward that ECE is independent of the reference probability Q. Thus, the selection of Q is only constrained by Equation 10. 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 4: the ellipse representing cross-entropy has always the same size. However, the inner small gray ellipse representing posterior entropy of Q may increase or decrease depending on the choice of the reference Q.

3.3. Choosing a reference Q for intuitive interpretation

The selection of the reference probability Q is constrained, because Equation 10 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. Considering that, in this paper 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 θ. This would only be achieved if the fact finder obtains the following posterior probabilities:

$$\begin{aligned} Q(\theta_p|E) &= 1\,, \quad \theta_p \text{ is true} \\ Q(\theta_p|E) &= 0\,, \quad \theta_d \text{ is true} \end{aligned} \quad (16)$$

which will be referred to as the *oracle* posterior distribution. If this oracle distribution is selected as a reference, 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 in order to know the true value of the hypotheses over many forensic cases. If the forensic system is 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.

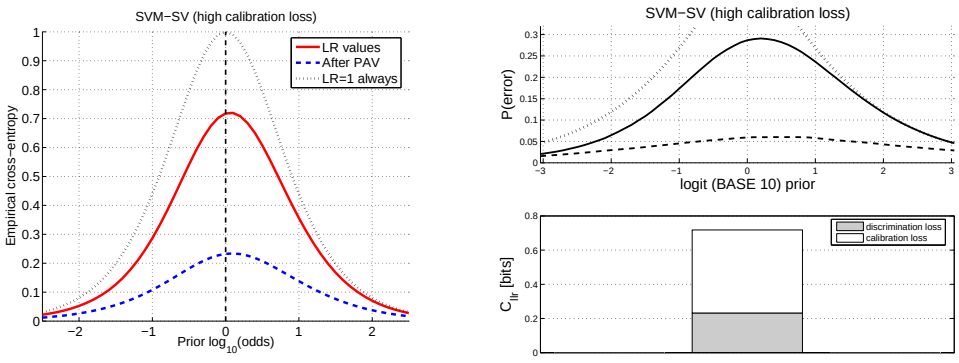
4. The ECE Plot

In this paper we propose to represent 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 $P(\theta_p|E)$ are computed using the LR values for the evaluation set and Equation 1. The value of ECE (Equation 15) is then represented as a function of $P(\theta_p)$.

Figure 5(a) shows an example of ECE plot for a sample ATVS-UAM system. The solid curve is the ECE (average information loss) of the LR values computed by the system. The higher this ECE curve, the higher the information needed on average in order to know the true hypothesis, 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 is the system which optimizes ECE while preserving discrimination [3][2]. The calibrated system is obtained from the forensic system using the Pool Adjacent Violators (PAV) algorithm (see [3] 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 neutral case is the prior, which is independent of the system.

[2]This system will be obtained from the forensic system, both having the same DET curve.



(a) (b)

Figure 5: Comparison of the ECE plot (a) to APE plots and C_{llr} (b). ATVS SVM-SV system. NIST SRE 2006 protocol.

Thus, according to Equation 9, the cross-entropy of the neutral system is simply the entropy of the prior probability, given by Equation 7. This neutral system plays an important role: if the ECE value of the forensic system is greater than the entropy of the neutral system, then the forensic system will lose 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.

4.1. Comparison to other performance measures

ECE and the already proposed C_{llr} are closely related. From Equations 6 and 15 it is straightforward that

$$C_{llr} = ECE|_{P(\theta_p)=0.5} \tag{17}$$

Thus, C_{llr} is a value which summarizes ECE. Moreover, ECE at a given prior represents the expected cost of taking decisions using any value of the decision costs, whose value at $P(\theta_p) = 0.5$ is C_{llr}. The interpretation of C_{llr} 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.

Another representation is proposed in [3], namely the APE plot, which represents the performance of a speaker detector in a wide range of applications in terms of error rates. If we set the $C_{fa} = C_{fr} = 1$, and we also assume that Bayes thresholds τ_B are used for taking decisions from the posterior $P(\theta_p|E)$, then Equation 3 represents the total error rate for τ_B, which is shown by the APE plots as a prior-dependent measure. It can be demonstrated [3] that the integral of the APE plot over the prior is proportional to C_{llr}. Therefore, reducing the error rate for τ_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 5. It is shown that the ECE value gives similar intuition about the calibration of the system as the APE plots. However, APE plots represents a given error rate due to decisions, but in ECE plots no decision is assumed to be taken. Also, it is clearly shown that the value of C_{llr} is the value of ECE at $P(\theta_p) = 0.5$[3].

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. NIST SRE 2006 protocol was followed in order to conduct the tests. An example of presenting the average information supplied by the forensic system in court based on a real case is also shown.

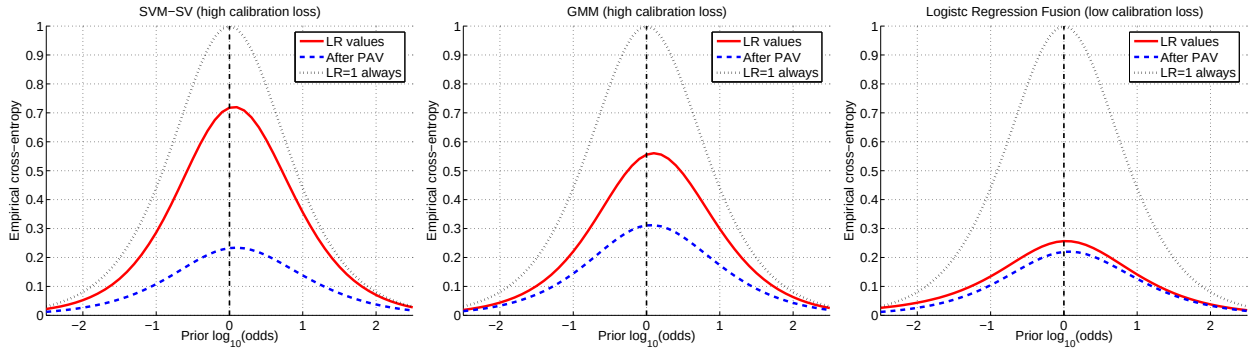
5.1. Database, evaluation protocol and systems

A forensic testing simulation has been performed using the evaluation protocol proposed in NIST 2006 SRE [13]. All the results presented in this paper correspond to the 1conv4w-1conv4w condition (608 speakers), where there is one conversation side for model training and one conversation side for testing. The length of the conversations is typically five minutes, with an average of 2.5 minutes after silence removal. For this condition, more than 50.000 score computations per system were performed. The database used in NIST SRE 2006 has been partially extracted from the MIXER corpus [13], but a significant amount of additional multi-channel and multi-language data was acquired in order to complete the corpus for the evaluation. It includes different communication channels, handsets, microphones and languages, and represents well the quality and diversity of real telephone conversations. Background data for training the system has been extracted from the NIST SRE 2005 database and protocol [13].

Two score-based systems have been used in order to obtain the scores from each recovered-control speech pair. On the one hand, a GMM-UBM-MAP system is used [17, 7]. On the other hand, we use a SVM-Supervector (SVM-SV) system, which is based on the classification of GMM mean-supervectors using support vector machines. Details can be found in [18, 7]. Nuisance Attribute Projection (NAP) technique has been used in order to compensate session variability [18]. It is important to notice that no LR computation technique has been used for reducing the calibration loss of the scores from the experiments conducted with the individual systems.

The two systems have been fused via logistic regression

[3]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_{llr}



(a) (b) (c)

Figure 6: ECE plots for the individual systems based on SVM-SV (a) and GMM (b), and for the fused system using logistic regression (c).

[19], a linear fusion where the transformation is trained in order to optimize an evaluation objective. In [19] it is demonstrated that, under some circumstances, such objective function is C_{llr}. Therefore, logistic regression not only fuses the scores coming from the individual systems, but it also tends to calibrate them. Logistic regression has been performed using the FoCal toolkit[4].

5.2. Information-theoretical evaluation

In Figure 6 the ECE plots are shown for the individual systems and for their fusion via logistic regression. Figures 6(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 a system and such 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.

Figure 6(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. Although $C_{llr} = ECE|_{P(\theta_p)=0.5}$ was used as an optimization objective, in this case ECE was reduced for every prior, because once a LR value is computed by logistic regression, it can be used for any other prior. Also, the difference between the dashed and solid curves is small, which means a small information loss due to a lack of calibration.

5.3. Presenting the average information supplied by the system in court

Imagine a scenario where the prosecutor presents a piece of evidence consisting of an incriminating questioned recording containing some utterances coming from one of 11 possible speakers. A suspect is appointed from police investigations, one of the 11 speakers, and some recordings are obtained from him. Considering only this background information, the fact finder may assign a prior probability $P(\theta_p) = \frac{1}{11}$ that θ_p is true (the suspect is the source of the questioned speech)[5]. The court gives the forensic speech scientist both recordings, and the fact finder also insists the scientist's analytical technique must comply with Daubert-like rules.

Taking into account all those elements, the forensic scientist uses one of the presented systems in order to compute a LR value to report the fact finder[6]. However, considering the admissibility requirements of Daubert rules, the foresic 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, the scientist includes the ECE plot of the forensic test in order to explain the fact finder the information given by the system in the inferential process.

If the fact finder so desires, the scientist may explain in court how the average information would be improved over many forensic cases by the use of the forensic system. Imagine that the scientist uses the fused system presented in Figure 6(c), which obtains a good value of ECE. Thus, the argument of the scientist should be as follows:

- Before knowing the weight of the evidence, and given that the prior probabilities have been set to $P(\theta_p) = 0.1$, the ECE plot shows that, using this system, we need 0.46 bits of information on average in order to know the true value of the hypothesis over cases like this one (dotted curve of Figure 6(c) at the prior odds $P(\theta_p)/P(\theta_d) = \frac{1}{10}$).
- After analyzing the weight of the evidence, more information has been obtained, and we will need only 0.12 bits on average in order to know the true value of the hypothesis over cases like this one (solid curve of Figure 6(c) at the prior odds $P(\theta_p)/P(\theta_d) = \frac{1}{10}$).
- If we had used the calibrated system, we would have need 0.1 bits on average in order to know the true value of the hypothesis (dashed curve of Figure 6(c) at the prior odds $P(\theta_p)/P(\theta_d) = \frac{1}{10}$). However, it has to be clear that this calibrated results are not feasible in practice, because the forensic scientist needs to know information

[4] FoCal is available at http://niko.brummer.googlepages.com/.

[5] For the presented simplified example, no other information is assumed to be present in the forensic case. However, in a real case the background information I may include more circumstantial information or other evidence sources.

[6] Many questions regarding the adequacy of the forensic testing database with respect to the real forensic field data may arise, as well as issues like population selection and reporting procedures. Such topics are out of the scope of this paper, but some discussion about them can be found in recent work from the authors [7].

about the true answers of the hypotheses in order to obtain this calibrated system.

6. Conclusions

In this paper, an analysis of the influence of the forensic speaker recognition system in the decision of a fact finder about a given forensic case has been presented in terms of information. Information theory has been used in order to derive empirical cross-entropy (ECE) as a measure of accuracy of a forensic speaker recognition system, according to other equivalent measures such as $U_{\log}$ or NCE. ECE can be interpreted as the average information needed by the fact finder over cases and after evidence analysis in order to know whether the recovered and control speech samples come from the same source or not. ECE considers the uncertainty about the hypotheses involved in a case in the light of the evidence and the rest of knowledge in the case. Moreover, it also measures the information loss due to a non-perfect LR calibration. This derivation has led to a novel elegant representation, the ECE plot, which allows presenting the average information supplied by evidence analysis in court with a clear separation of roles. The derived representation allows the transparent reporting of the performance of the system in terms of such information-theoretical magnitudes. The proposed ECE plot has been compared to other assessment methods such as APE plots and C_{llr}. As a conclusion, the authors believe that the proposed information-theoretical interpretation may be easy to understand by fact finders, aiding decisions about admissibility according to Daubert rules and other similar criteria.

Another advantage of the presented technique is its adequacy to other forensic disciplines where LR values are used for evidence evaluation, such as glass and paint analysis [20].

7. Acknowledgments

This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. The authors want to thank Niko Brümmer, Colin Aitken and Grzegorz Zadora for fruitful comments, suggestions and discussions. The authors also thank one of the reviewers for extensive comments which have greatly improved the quality of the final paper.

8. References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd ed.*, Wiley Interscience, 2006.

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

[4] 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.

[5] A. Drygajlo, "Forensic automatic speaker recognition [exploratory DSP]," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 132–135, 2007.

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

[7] 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.

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

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

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

[11] 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.

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

[13] 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.

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

[15] N. Brümmer, "Application-independent evaluation of speaker detection," in *Proc. of Odyssey*, 2004, pp. 33–40.

[16] NIST, "A tutorial introduction to the ideas behind normalized cross-entropy and the information-theoretic idea of entropy," Tech. Rep., 2004, Available at http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/NCE.pdf.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[18] W. M. Campbell, D. E. Sturim, D. A. Reynols, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. of ICASSP*, Toulouse, France, 2006, pp. 97–100.

[19] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim, "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.

[20] 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 (in IAS 2007)*, 2007, pp. 411–416.