Problems of Forensic Sciences 2007, LXIX, 62–70

ISSN 1230-7483

INFORMATION-THEORETICAL EVALUATION OF LIKELIHOOD RATIOS

Daniel RAMOS, Joaquin GONZALEZ-RODRIGUEZ

ATVS – Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain

Abstract

In this paper, we present an evaluation methodology for any likelihood-ratio *LR*-based forensic discipline. The proposed method is based on information theory, showing the performance of the system in the form of cross-entropy as a measure of uncertainty and deviation between probability distributions. This assessment methodology takes into account not only the discrimination of the technique in use, but also the calibration of *LR* values. In order to illustrate the proposed method, an example is shown using an *LR*-based speaker recognition system, simulating a blind forensic test following the NIST Speaker Recognition Evaluation 2006 protocol.

Key words

Cross-entropy; Information theory; Likelihood ratio; Assessment.

Received 16 July 2007; accepted 12 September 2007

1. Introduction

In recent years, interest in the classical debate about presentation of forensic evidence in a court of law has significantly increased [12], motivated by several reasons. First, the establishment of the American Daubert rules for the admissibility of scientific evidence in trials [15] has led to the questioning of the scientific value of some well-established techniques. Second, partly due to some critical errors in positive identification reports (see, e.g. [7]), the accuracy of existing techniques which up to now have been assumed by courts as error-free are starting to be questioned. In order to cope with the emerging requirements, standardisation and forensic testing should be key points in the presentation of the accuracy of the systems (in courts) in an acceptable way. In this sense, a likelihood ratio (*LR*) approach to evidence analysis [1, 6] has been proposed as a common framework for forensic interpretation of evidence, emulating procedures in DNA. In this *LR* framework, the decision about source attribution is performed by the fact finder, who also defines the hypotheses in the case, considering the prior probabilities of the hypotheses and the *LR* computed by the forensic scientist [1].

One of the main advantages of *LR*-based methods is their testability. Opinions about hypotheses are expressed by fact finders in the form of posterior probabilities. Therefore, there is a need to measure not only the discrimination capabilities of the system, but how these probabilities affect the correctness of the decisions. Highly discriminant systems may lead to wrong posterior probabilities if they do not elicit calibrated confidences [2, 5].

In this paper, a methodology for the assessment of calibration effects in *LR*-based forensic evidence evaluation is presented. We present cross-entropy as a measure of the discrimination and calibration of the

probabilities obtained with the *LR* values computed by the scientist. An experimental example is shown in Section 4, where we simulate a comparative forensic testing of several robust approaches proposed in the literature for *LR*-based forensic speaker recognition systems [11]. Finally, conclusions are drawn in Section 5.

2. The likelihood ratio methodology

The *LR* framework for interpretation of evidence [1, 6] represents a mathematical and logical tool which presents many advantages in the forensic context. The objective is to compute the likelihood ratio (*LR*) as the degree of support for one hypothesis versus its opposite. We assume that evidence *E* involves comparison of a questioned trace recovered from the scene of a crime (e.g. a wire-tapped recording, a signature on a document, etc.) with some material from a known source, which may be a suspect (e.g. a recording from the suspect in controlled situations, a signature acquired from the suspect, etc.). Bayes' theorem states that:

$$\frac{Pr\ H_p|E,I}{Pr\ H_d|E,I} \quad LR \quad \frac{Pr\ H_p|I}{Pr\ H_d|I};$$

$$LR \quad \frac{f\ E|H_p,I}{f\ E|H_d,I}, \qquad \{1\}$$

where H_p (the suspect is the source of the recovered sample) and H_d (another individual is the source of the recovered sample) are typically the relevant hypotheses and I is the background information available in the case. The hypotheses are defined in the court from I, the prosecutor and defence propositions and often because of the adversarial nature of the criminal system.

3. Information-theoretical assessment of *LR* values

3.1. Uncertainty and information

In this section we present a measure of performance based on information theory. The *LR* values determined by the forensic scientist (referred to hereafter as the "scientist's" values) are compared with the *LR* values determined from the true values of the propositions (same-source, H_p or different-source, H_d), which are referred to as the "evaluator's" *LR* values. In an information-theoretical framework [4, 13], the uncertainty of a random variable (denoted H with two possible values $\{H_p, H_d\}$), given the evidence E, is measured by the conditional entropy[1], defined as:

$$U_{Pr}\ H|E$$

$$f\ e,H_i\ \log_2 Pr\ H_i|e\ de, \qquad \{2\}$$

$$e\ i\ p,d$$

where P_r denotes probability and f denotes probability density function (*pdf*). However, solution of {2} is not possible in general, since the integral may be difficult to evaluate. The problem is solved by comparing the scientist's *LR* values with the evaluator's *LR* values as follows. The evaluator knows the true states of H for each comparison. Thus, the corresponding *LR* values are for a same-source experiment and 0 for a different-source experiment. The corresponding posterior probabilities are 1 and 0 for H_p and H_d, respectively, using the odds form of Bayes Theorem. Also, the evaluator is perfectly reliable, because no error is introduced. A comparison of the scientist's *LR* values with the reliable evaluator's *LR* values is given by the conditional cross-entropy:

$$U_{\tilde{P}r|Pr}\ H|E \qquad f\ E \quad U_{\tilde{P}r|Pr}\ H|E \quad e$$

$$\tilde{P}r\ H_i \quad \tilde{f}\ e|H_i\ \log_2 Pr\ H_i|e\ de, \quad \{3\}$$

$$i\ p,d \qquad\qquad e$$

where a tilde indicates the probabilities computed by the evaluator. It can be demonstrated that assigning $LR =$ to all $e\ p$ (same-source evidence values) and $LR = 0$ to all $e\ d$ (different-source evidence values) gives:

$$U_{\tilde{P}r|Pr}\ H|E \quad \sim$$

$$\tilde{P}r\ H_i \quad \frac{1}{N_i\ e_j\ E_i}\ \log_2 Pr\ H_i|e_j \quad , \{4\}$$

$$i\ p,d$$

where N_p is the number of same-source comparisons and N_d is the number of different-source comparisons. Thus, {4} measures the conditional cross-entropy (or simply cross-entropy) between the evaluator's *LR* values and the scientist's *LR* values. For this reason, it is proposed as the metric formula for evaluation of the performance of an *LR* estimation procedure. The cross-entropy values in {4} depend on the prior probabilities of H since:

[1] We use the notation U (uncertainty) for entropy for clarity with respect to the notation for propositions (H_p, H_d).

$$Pr\ H_p|E \quad \frac{LR\ \frac{Pr\ H_p}{Pr\ H_d}}{1\quad LR\ \frac{Pr\ H_p}{Pr\ H_d}}. \qquad \{5\}$$

3.2. Discrimination and calibration

The ability of an *LR* procedure to discriminate is assessed by its performance in distinguishing whether a single pair of two samples comes from the same source or from different sources. For pairs selected from a database, it is known whether they come from the same source or from a different source. The performance of an *LR* procedure may be assessed by counting the number of false assignments using *LR* for a given threshold θ. In general, in decision theory, θ is determined from $Pr(H_p)$ and the utilities involved in the decision process [14]. Thus, a false positive and false negative rate for every value of θ may be defined. The discrimination performance of an *LR* procedure, defined by a set of *LR* values, is itself defined as the relationship between false positives and false negatives for every value of the threshold θ. Typical representations of discrimination performances are ROC (Receiver Operating Characteristic Curves) or DET (Detection Error Tradeoff) curves, the former of which plots correct detection rate versus false negatives and the latter of which plots false positives versus false negatives [8]. An example of a DET curve may be found in Figure 1. Two different sets of *LR* values have the same discrimination performance if, for every possible threshold θ in the first set of *LR* values, a threshold θ' can be found for the second set of *LR* values which presents the same false positive and false negative rates as the first set. It is derived from his definition that a monotonic transformation does not change the discrimination performance of an *LR* procedure. On the other hand, two different *LR* procedures may have a very different performance in terms of cross-entropy even if they have the same discrimination performance. This is due to a different calibration of each *LR* procedure, as defined in [5]. Cross-entropy measures the mean deviation of $Pr(H|E)$, from {5}, with respect to $\widetilde{Pr}(H|E)$, which is the perfectly reliable evaluator's posterior distribution. Calibration is used to compare two or more *LR* procedures; it measures the deviation of $Pr(H|E)$ with respect to $\widetilde{Pr}(H|E)$ for *LR* procedures for which the discrimination is the same. Some illustrating examples of the effects of a lack of calibration may be found in [2, 10, 11].

It is desirable to find a monotonic transformation which minimizes the value of the cross-entropy for a given *LR* procedure. This will provide a new set of transformed *LR* values which will be optimal in terms of cross-entropy constrained to not changing the discrimination performance of the original set of scientist's *LR* values. This new set of *LR* values will be calibrated in the sense that the cross-entropy cannot be reduced for the given discrimination performance. This new set of transformed *LR* values is called the calibrated *LR* set. Isotonic regression using the Pool Adjacent Violators (PAV) algorithm [2] is used to obtain the monotonic transformation which leads to the calibrated *LR* values. Details about the PAV algorithm may be found in [2]. However, the forensic scientist cannot use PAV in order to obtain the calibrated *LR* set, because PAV requires the true values of *H*, which are unknown by the forensic scientist.

3.3. The information plot

The measurements of discrimination and calibration described above allow cross-entropy to be represented as a function of $Pr(H_p)$ in a so-called information plot. Figure 2 shows several examples of information plots. The cross-entropy is represented for three cases. First, the solid curve is the cross-entropy of the forensic system *LR* values. Second, the dashed curve is the cross-entropy of the calibrated *LR* values obtained with the PAV algorithm [2]. Finally, the dotted curve is the cross-entropy of a neutral *LR* set, i.e., the *LR* is always 1. The reference curve does not change between *LR* procedures.

As can be seen in [4], if we use the evaluator's *LR* values, the evidence will give total certainty about the hypotheses and therefore the information gain would have been optimal. This is impossible to achieve, as the scientist does not have knowledge about the true answers. Thus, as the opinion of the scientist diverges from that of the evaluator, the cross-entropy (solid curve) increases, and the effective information about the propositions that is extracted from the evidence is lower. Moreover, the dashed curve shows the potential information that the system is able to deliver if it has been properly calibrated.

4. Case of study: forensic speaker recognition

In this section, we present an experimental example using different *LR* computation techniques with a speaker recognition system. The system used for *LR* computation is based on the classification of Gaussian

Mixture Model (GMM) mean supervectors using Support Vector Machines (SVM). Nuisance Attribute Projection (NAP) is used for session variability compensation. Thus, *LR* values have been obtained using scores from the ATVS – Biometric Recognition Group GMM-SVM-NAP speaker recognition system, as described in [11]. The system is based on the classification of GMM mean-supervectors using support vector machines. Details may be found in [3]. The comparative results presented here relate to two techniques for the evaluation of forensic evidence found in the literature, namely: 1) suspect-independent *LR* computation and 2) suspect-adapted *Maximum a Posteriori* (MAP) *LR* computation. In suspect-independent within-source estimation, 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. On the other hand, suspect-adapted MAP estimation of within-source distributions adapts the global distribution to the suspect distribution obtained from suspect speech samples obtained at the trial. Therefore, an adapted within-source pdf is obtained. See [11] for details.

Experiments have been performed using the evaluation protocol proposed in the 2006 Speaker Recognition Evaluation by the American National Institute of Standards and Technology (NIST 2006 SRE). These evaluations have constituted a common framework for the development of automatic speaker recognition technology, with protocols and databases which are considered standard. See [9] for details. Using this experimental protocol, more than 5000 same-source and 45000 different-source *LR* values have been obtained.

In Figures 1 and 2 we compare the performance of the different evaluated techniques. Results are presented in the form of DET curves and information plots. In Figure 1 we can see the DET plots showing the discrimination performance of the different evaluated *LR* computation techniques using the GMM-SVM-NAP speaker recognition system. It can be observed that the discrimination performance is better for the suspect-adapted case. Thus, suspect-adapted *LR* computation exploits suspect specificities in order to lead to a better discrimination performance.

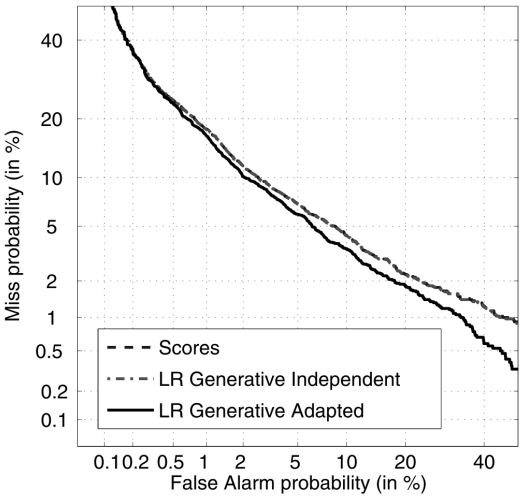

Fig. 1. DET curves comparing speaker recognition scores and described *LR* procedures. NIST SRE 2006 database and protocol.

Figures 2 shows the proposed information-theoretical methodology for presenting forensic system testing results. It is observed that the discrimination performance (dashed curve) is better for the suspect-adapted case, as was shown in 1. The calibration per-

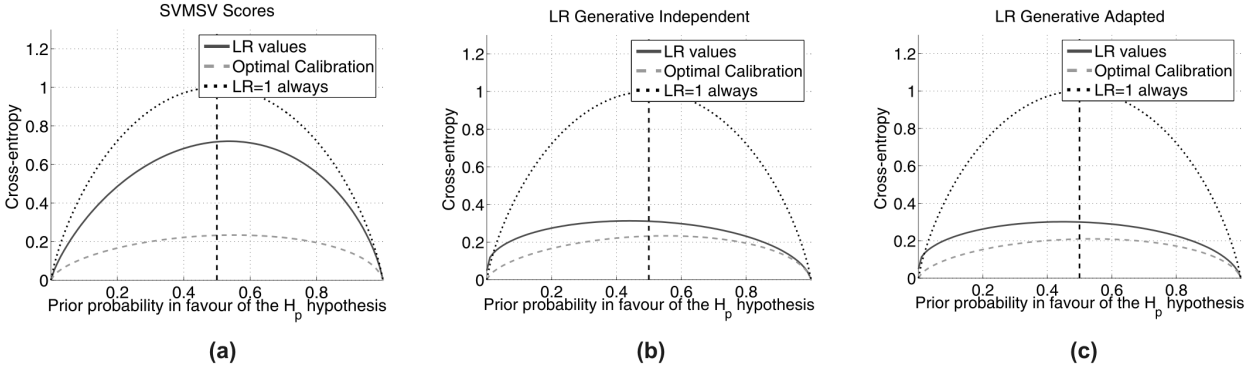

Fig. 2. Information plots comparing speaker recognition scores and the described *LR* procedures. NIST SRE 2006 database and protocol.

formance is very similar for both *LR* computation techniques, significantly improving the scores.

5. Conclusions

In this paper a methodology of assessment of *LR* values based on information theory has been presented. The proposed technique measures the loss of information due to the deviation of the forensic scientist's *LR* values from the *LR* values computed with knowledge of the true answers. This method measures not only the discrimination capabilities of the generated *LR* values, but also their lack of calibration, which imply an information loss. An example using *LR* based speaker recognition technique is presented, where the proposed evaluation techniques are compared to discrimination measures as DET plots using a clear and standard protocol such as those developed by NIST in their 2006 SRE. This is an example of a methodology for evaluating the performance of a set of *LR* values in a controlled and transparent way.

Acknowledgements
This work has been supported by the Spanish Ministry of Sciences and Technology under project TEC2006-13170-C02-01. The authors wish to thank Prof. Colin Aitken, School of Mathematics, University of Edinburgh, UK and Dr. Grzegorz Zadora, Institute of Forensic Research, Krakow, Poland for useful comments and discussion.

References

1. Aitken C. G. G., Taroni F., Statistics and the evaluation of evidence for forensic scientists, John Wiley & Sons, Chichester 2004.
2. Brümmer N., du Preez J., Application independent evaluation of speaker detection, *Computer Speech and Language* 2006, 20, 230–275.
3. Campbell W. M., Sturim D. E., Reynolds D. A., Support vector machines using GMM supervectors for speaker verification, *Signal Processing Letters* 2006, 13, 308–311.
4. Cover T. M., Thomas J. A., Elements of information theory, Wiley Interscience, Chichester 2006.
5. de Groot M. H., Fienberg S. E., The comparison and evaluation of forecasters, *The statistician* 1982, 32, 12–22.
6. Evett I. W., Towards a uniform framework for reporting opinions in forensic science casework, *Science and Justice* 1998, 38, 198–202.
7. Heath D., Bemton H., Portland lawyer released in probe of Spanish bombings, Seattle Times, May 21, 2004; http://www.law.asu.edu/?id=8857.
8. Martin A. [et al.], The DET curve in assessment of decision task performance, Proceedings of European Conference on Speech Communications and Technologies, EuroSpeech, Rhodes (Greece) 1997.
9. NIST speech group website; http://www.nist.gov/speech.
10. Ramos D., Gonzalez-Rodriguez J., Zadora G. [et al.], Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation, Proceedings of International Workshop of Computational Biometrics, Manchester 2007 [in press].
11. Ramos-Castro D., Gonzalez-Rodriguez J., Ortega-Garcia J., Likelihood ratio calibration in transparent and testable forensic speaker recognition, Proceedings of Odyssey, San Juan (Puerto Rico), 2006.
12. Saks M. J., Koehler J. J., The coming paradigm shift in forensic identification science, *Science* 2005, 309, 892–895.
13. Shannon C. E., A mathematical theory of communication, *Bell System Technical Journal* 1948, 27, 379–423, 623–656.
14. Taroni F., Bozza S., Aitken C. G. G., Decision analysis in forensic science, *Journal of Forensic Sciences* 2005, 50, 894–905.
15. *U.S. Supreme Court, Daubert v. Merrel Dow Pharmaceuticals* 2003 [509 U.S. 579].

Corresponding author
Daniel Ramos
Universidad Autonoma de Madrid
C/. Francisco Tomas y Valiente 11
E-28049 Madrid
e-mail: daniel.ramos@uam.es

Problems of Forensic Sciences 2007, LXIX, 62–70

OCENA JAKOŚCI MODELU DO OBLICZANIA WARTOŚCI ILORAZU WIARYGODNOŚCI ZA POMOCĄ METODY OPARTEJ NA TEORII INFORMACJI

1. Wstęp

W ostatnich latach zainteresowanie od dawna już poruszanym zagadnieniem formy prezentacji dowodu naukowego w sądzie znacznie wzrosło [12], co spowodowane zostało kilkoma przyczynami. Po pierwsze, ustanowiono tzw. reguły Dauberta mówiące, kiedy dowód naukowy powinien być brany pod uwagę w procesie sądowym [15], co doprowadziło do pytania o naukowość części z powszechnie stosowanych technik analizy materiału dowodowego. Po drugie, z powodu kilku poważnych błędów (w rodzaju odpowiedzi fałszywie pozytywnej) w ocenie materiału dowodowego dokonanego przez biegłych (np. [7]) zachwiane zostało poczucie, że stosowane dla potrzeb wymiaru sprawiedliwości techniki są precyzyjne, zwłaszcza że były one dotychczas uważane za bezbłędne. W celu sprostania narastającym wymaganiom, standaryzacja oraz testowanie wspomnianych metod powinny stać się kluczowymi zadaniami pozwalającymi na zaprezentowanie ich poprawności w sądzie w powszechnie akceptowalny sposób. W tym sensie obliczanie ilorazu wiarygodności (*LR*) w celu oceny wartości dowodowej [1, 6] zostało zaproponowane jako uniwersalny sposób do interpretacji dowodów naukowych na przykład w analizie wartości dowodowej profilu DNA. W metodzie *LR* decyzja o przedmiocie prowadzonej analizy jest dokonywana przez osobę prowadzącą dochodzenie lub sędziego, który definiuje hipotezy, rozważa prawdopodobieństwa *a priori* tych hipotez i ocenia wartość *LR* obliczoną przez biegłego sądowego [1].

Jedną z pozytywnych cech oceny wartości dowodowej bazującej na obliczaniu wartości *LR* jest to, że można je testować. Jak wspomniano, opinie o hipotezach są wyrażane przez osobę prowadzącą dochodzenie w formie prawdopodobieństw *a posteriori*. Dlatego też konieczne jest mierzenie nie tylko właściwości dyskryminacyjnych metod, ale również ocena tego, jak uzyskane wartości *LR* wpływają na poprawność podejmowanych decyzji na podstawie prawdopodobieństwa *a posteriori*. System ten, posiadający wysoce dyskryminujące właściwości, może jednocześnie doprowadzać do otrzymania niepoprawnych wartości prawdopodobieństw *a posteriori*, jeżeli nie zastosuje się procedury kalibracji [2, 5]. Przedstawiona w niniejszym artykule metoda oceny modeli *LR* analizuje też wpływ procesu kalibracji. Ponadto omówiono entropię krzyżową jako miarę dyskryminacji i kalibracji prawdopodobieństw *a posteriori* uzyskanych z wartości *LR* obliczonych przez biegłego sądowego. Przykład praktyczny przedstawiono w rozdziale 4, gdzie pokazano rezultaty testu porównawczego kilku rzetelnych metod proponowanych w literaturze przedmiotu do obliczania wartości *LR* w systemach rozpoznawania mowy [11].

2. Metodologia ilorazu wiarygodności

Oszacowanie wartości dowodowej poprzez obliczenie ilorazu wiarygodności [1, 6] jest narzędziem matematycznym i logicznym, które – z punktu widzenia wymiaru sprawiedliwości – posiada wiele pozytywnych cech. Istotne jest bowiem obliczanie ilorazu wiarygodności (*LR*) jako siły wsparcia dla jednej z hipotez będącej w opozycji do innej hipotezy. Zakłada się, że dowód E jest związany z problemem porównania śladu ujawnionego na miejscu zdarzenia (np. nagranie, podpis pod dokumentem itd.) z materiałem porównawczym zabezpieczonym np. od podejrzanego (nagranie głosu pobrane od podejrzanego w kontrolowanych warunkach, podpisy zebrane od podejrzanego itp.). Z teorii Bayesa wynika, że:

$$\frac{Pr\ H_p|E,I}{Pr\ H_d|E,I} \quad LR \quad \frac{Pr\ H_p|I}{Pr\ H_d|I};$$

$$LR \quad \frac{f\ E|H_p,I}{f\ E|H_d,I}, \qquad \{1\},$$

gdzie H_p (podejrzany jest źródłem próbki dowodowej) i H_d (inna osoba jest źródłem próbki dowodowej) są typowymi przykładami rozpatrywanych hipotez, a I oznacza informację podstawową o przestępstwie dostępną w danej sprawie. Hipotezy te są definiowane przez sąd na podstawie wiedzy o I, a hipotezy prokuratora i obrony są przeciwstawne z natury samego procesu sądowego.

3. Oszacowanie modelu *LR* za pomocą metody opartej na teorii informacji

3.1. Niepewność

W tym rozdziale omówiona zostanie ocena modelu *LR*, która oparta jest na teorii informacji. Wartości *LR* wyznaczone przez biegłego (nazywane później „wartościami naukowca") są porównywane z wartościami *LR* wyznaczonymi jako prawdziwe wartości analizowanych hipotez (to samo źródło H_p lub różne źródła H_d), które są

dalej nazywane „wartościami egzaminatora". W teorii informacji niepewność zmiennej losowej (oznaczonej jako H z dwoma możliwymi wartościami $\{H_p, H_d\}$) pod warunkiem, że rozważany jest dowód E, mierzona jest jako entropia[1] warunkowa zdefiniowana jako:

$$U_{Pr}\ H|E \quad f\ e, H_i\ \log_2 \Pr\ H_i|e\ de, \qquad \{2\}$$

gdzie Pr oznacza prawdopodobieństwo, a f funkcję gęstości prawdopodobieństwa. Ponieważ całka w równaniu {2} nie ma rozwiązania, to równanie również go nie posiada. Dlatego też problem rozwiązywany jest poprzez porównanie wartości LR wyznaczonego przez „naukowca" z wartością uzyskaną przez „egzaminatora", który zna prawdziwą wartość stanów H dla każdego porównania. Tak więc „egzaminator" przypisuje wartość wszystkim wartościom LR dla porównań typu „to samo źródło próbek" i 0 dla porównań typu „różne źródło próbek". Odpowiadające im wartości prawdopodobieństw *a posteriori* są odpowiednio równe 1 i 0 dla H_p i H_d. Ponadto wartości przypisane przez „egzaminatora" są absolutnie wiarygodne, ponieważ żaden błąd nie jest z nimi związany. Porównanie wartości LR „naukowca" z wartością uzyskaną przez wiarygodnego „egzaminatora" może być wyrażona poprzez względną entropię krzyżową:

$$U_{\tilde{P}r|Pr}\ H|E \quad _{f\ E}\ U_{\tilde{P}r|Pr}\ H|E \quad e$$

$$\tilde{P}r\ H_i \quad \tilde{f}\ e|H_i \quad \log_2 Pr\ H_i|e\ de, \{3\}$$

gdzie tylda oznacza prawdopodobieństwa obliczone przez osobę „egzaminatora". Można wykazać, że przypisując $LR =$ do wszystkich e E_p (wartości uzyskane w przypadku, gdy porównywane próbki pochodzą z tych samych źródeł) i $LR = 0$ dla wszystkich e E_d (wartości uzyskane w przypadku, gdy porównywane próbki pochodzą z różnych źródeł) daje:

$$U_{\tilde{P}r|Pr}\ H|E \quad \sim$$

$$\tilde{P}r\ H_i \quad \frac{1}{N_i\ e_j\ E_i} \quad \log_2 Pr\ H_i|e_j \quad , \quad \{4\}$$

gdzie N_p jest liczbą porównań dokonanych pomiędzy próbkami pochodzącymi z tego samego obiektu, a N_d jest liczbą porównań dokonanych pomiędzy próbkami pochodzącymi z różnych obiektów. Dlatego też równanie {4} mierzy względną entropię krzyżową (lub w skrócie entropię krzyżową) pomiędzy wartością LR „egzaminatora" oraz wartością LR „naukowca" i została zaproponowana jako miara jakości modelu służącego do wyznaczania LR. Ponadto wartość entropii krzyżowej w równaniu {4} zależy od wartości prawdopodobieństwa *a priori* dla H:

$$Pr\ H_p|E \quad \frac{LR\ \frac{Pr\ H_p}{Pr\ H_d}}{1 \quad LR\ \frac{Pr\ H_p}{Pr\ H_d}}. \qquad \{5\}$$

3.2. Dyskryminacja i kalibracja

Zdolność dyskryminacji danego modelu LR zdefiniowana jest jako zdolność do rozróżniania porównywanych par próbek pochodzących z tego samego źródła czy też z różnych źródeł. W przypadku próbek istniejących w bazie danych wiadome jest, czy pochodzą one z tej samej próbki (źródeł), czy z różnych próbek (źródeł). Efektywność modelu LR może być ustalona przez obliczenie liczby fałszywych oszacowań przy założeniu konkretnego punktu odniesienia θ, tj. konkretnej wartości LR. Generalnie rzecz biorąc, w teorii podejmowania decyzji θ jest wyznaczane z $Pr(H_p)$ i funkcji użyteczności, która jest częścią procesu decyzyjnego [14]. Dlatego też liczba odpowiedzi fałszywie pozytywnych i fałszywie negatywnych dla każdej wartości θ może zostać zdefiniowana. Sposób, w jaki działa czynnik dyskryminujący, czyli analizowany model LR, jest definiowany przez związek pomiędzy odpowiedziami fałszywie pozytywnymi i fałszywie negatywnymi uzyskanymi dla każdej wartości θ. Typową metodą przedstawiania siły dyskryminującej modelu są krzywe ROC (ang. receiver operating characteristics – wykres charakterystyki uzyskiwanych wyników), w której procent odpowiedzi poprawnych zobrazowany jest w funkcji procenta odpowiedzi fałszywie negatywnych oraz krzywe DET (ang. detection error tradeoff – wykres odpowiedzi fałszywych) obrazujące procent odpowiedzi fałszywie pozytywnych względem procenta odpowiedzi fałszywie negatywnych [8]. Przykład krzywej DET zaprezentowano na rycinie 1. Dwie różne procedury obliczania wartości LR mają tę samą wartość dyskryminacyjną, jeżeli dla każdej z możliwych wartości θ w pierwszym zbiorze wartości LR można znaleźć wartość θ' w drugim zbiorze wartości LR, która prezentuje ten sam procent odpowiedzi fałszywie pozytywnych i fałszywie negatywnych. Bardziej formalnie można to wyrazić w ten sposób, że transformacja monotoniczna nie zmienia właściwości dyskryminacyjnych analizowanego modelu LR. Z drugiej strony dwa różne modele obliczania LR mogą dawać bardzo różne wyniki z punktu widzenia entropii krzyżowej, mimo iż ich właściwości dyskryminacyjne są identyczne. Związane jest to z ich różną kalibracją [5]. Entropia krzyżowa mierzy średnie odchylenie $Pr(H|E)$ z {5}, w odniesieniu do wyrażenia

[1] Symbol U (niepewność) został użyty do oznaczenia entropii w związku z hipotezą (H_p, H_d), aby uczynić notację bardziej przejrzystą.

$\widetilde{Pr}(H|E)$, które jest rozkładem *a posteriori* „egzaminatora". Kalibracja jest stosowna w celu porównania dwóch lub więcej modeli *LR* i mierzy ona odchylenie $Pr(H|E)$ w stosunku do $\widetilde{Pr}(H|E)$ modeli, dla których właściwości dyskryminacyjne są takie same.

W kalibracji pożądane jest znalezienie monotonicznej transformacji, która minimalizuje wartość entropii krzyżowej dla danego modelu *LR*. Jako efekt uzyskuje się zbiór nowych (po transformacji) wartości *LR*, który będzie optymalny w sensie znaczenia entropii krzyżowej obliczanej na założeniach, że niezmienna jest siła dyskryminacyjna wyjściowego (oryginalnego) zbioru wartości *LR* wyznaczonych przez „naukowca". Ten nowy zbiór wartości *LR* jest kalibrowany w tym sensie, że entropia krzyżowa nie może być już bardziej zredukowana dla danego modelu dyskryminacyjnego. Nowe wartości *LR* uzyskane po transformacji zwane są zbiorem skalibrowanych wartości *LR*. Izotoniczna regresja z zastosowaniem algorytmu PAV (ang. pool adjacent violators) [2] może być zastosowana w celu przeprowadzenia transformacji, która prowadzi do uzyskania skalibrowanych wartości *LR*. Szczegóły o algorytmie PAV można znaleźć w literaturze przedmiotu [2]. Niemniej jednak biegły sądowy nie może stosować algorytmu PAV w celu uzyskania skalibrowanego zbioru wartości *LR*, ponieważ PAV wymaga znajomości prawdziwej wartości *H*, która nie jest znana biegłemu sądowemu.

3.3. Wykres informacyjny

Opisany powyżej pomiar siły dyskryminacji procesu kalibracji pozwala ukazać entropię krzyżową jako funkcję $Pr(H_p)$ na tzw. wykresie informacji. Rycina 2 przedstawia kilka przykładów wykresów informacji, na których entropia krzyżowa uwidacznia się w trzech przypadkach. Linia ciągła to entropia krzyżowa wartości *LR* uzyskanych przez biegłego. Linia kreskowana to entropia krzyżowa wartości *LR* uzyskanych po kalibracji, w której zastosowano algorytm PAV [2]. Linia kropkowana to entropia krzyżowa zbioru neutralnych wartości *LR*, tj. takich, gdy *LR* jest zawsze równe 1. Jest to tzw. krzywa odniesienia, która jest stała dla różnych modeli obliczania wartości *LR*.

Jak wykazano w literaturze przedmiotu [4], jeżeli stosujemy wartości *LR* „egzaminatora", to wówczas dowód będzie dawał całkowitą pewność o hipotezie (tzn. zdarzenie opisane przez *H* jest zdarzeniem pewnym) i dlatego też wzrost liczby informacji powinien być optymalny. Sytuacja taka nie jest możliwa, ponieważ biegły nigdy nie ma wiedzy o prawdziwym stanie analizowanej sytuacji. Dlatego też opinia biegłego („naukowca") różni się od opinii oceniającego („egzaminatora") tym, że na wykresie informacji entropia krzyżowa (linia ciągła) znajduje się powyżej linii opisującej efektywną informację o hipotezach, która jest uzyskiwana z danych o dowodzie. Ponadto linia przerywana ilustruje potencjalną informację, którą system może dostarczyć, jeżeli jest on poprawnie skalibrowany.

4. Przykład – rozpoznawanie mówcy

W niniejszym rozdziale zaprezentowano przykład oceny różnych technik obliczania *LR* w systemach rozpoznawania mówcy. Model używany do obliczeń *LR* oparty jest na klasyfikacji z zastosowaniem klasyfikatorów maksymalnoodległościowych (ang. support vector machines) uwzględniający gaussowski model mieszany (ang. Gaussian mixture model) oraz model NAP (ang. nuissance attribute projection) do kompensacji zmienności (wariancji). Model GMM-SVM-NAP opracowano w Sekcji Biometryki działającej na Uniwersytecie Autonomicznym w Madrycie [11]. Szczegółowe informacje można również znaleźć w literaturze przedmiotu [3]. Rezultaty przedstawione w tym miejscu dotyczą dwóch technik stosowanych do oceny materiału dowodowego, które można znaleźć w literaturze, tj.: 1) obliczanie wartości *LR* przy założeniu „niezależnego podejrzanego", 2) obliczanie wartości *LR* przy założeniu „zaadoptowanego podejrzanego" (ang. suspect-adapted), stosując technikę uzyskania maksymalnej wartości prawdopodobieństw *a posteriori* (*maximum a posteriori* – MAP). W przypadku obliczeń przy założeniu „niezależnego podejrzanego" proponuje się w celu obliczenia wariancji wewnątrzobiektowej technikę, w której zakłada się, że dokładny model rozkładu wariancji wewnątrzobiektowej dla konkretnego podejrzanego można uzyskać poprzez analizę próbek pobranych w tych samych warunkach pomiarowych w trakcie analizy różnych osób. Z kolei w modelu opisanym w punkcie 2) oszacowanie rozkładu zmienności wewnątrzobiektowej polega na zaadaptowaniu rozkładu gęstości prawdopodobieństwa, który uzyskuje się z próbek mowy podejrzanego zebranych w trakcie prowadzonego dochodzenia [11].

Eksperyment wykonano, stosując protokół zaproponowany i opisany w ocenie systemów rozpoznawania mówcy przez Amerykański Narodowy Instytut Standardów i Technologii (NIST 2006 SRE). Oceny te przyczyniły się do rozwoju automatycznej procedury rozpoznawania mówcy, na którą składa się protokół obliczeniowy oraz baza danych. Obecnie procedura ta jest używana standardowo [9]. Stosując ten sposób postępowania, wyznaczono ponad 5000 wartości *LR* w przypadku porównań wewnątrzobiektowych i 45 000 wartości *LR* w przypadku porównań pomiędzyobiektowych.

Na rycinie 1 i 2 porównano rezultaty uzyskane za pomocą obu technik. Rezultaty są przedstawione w formie krzywych DET i krzywych informacyjnych. Na rycinie 1 można zobaczyć, że krzywe DET pokazują zdolności dyskryminujące ocenianych technik obliczania *LR* po za-

stosowaniu systemu GMM-SVM-NAP rozpoznawania mówcy. Można zaobserwować, że właściwości dyskryminujące są lepsze dla modelu z założeniem „zaadoptowany podejrzany", co spowodowane jest faktem, że podejście to wykorzystuje parametry charakteryzujące podejrzanego.

Rycina 2 pokazuje rezultaty proponowanej metodyki opartej na teorii informacji w przypadku testowania konkretnego modelu obliczania *LR*, a zastosowanego dla potrzeb wymiaru sprawiedliwości. Można zaobserwować, że właściwości dyskryminujące (linia kreskowana) są lepsze, gdy zastosowano założenie „zaadoptowany podejrzany", co też pokazuje rycina 1. Wynik kalibracji jest bardzo podobny dla obu technik obliczania wartości *LR* i znacząco poprawia uzyskiwane wyniki.

5. Wnioski

W artykule zaprezentowano metodykę opartą na teorii informacji, a służącą do oceny modeli stosowanych do obliczania wartości *LR*. Proponowany sposób postępowania pozwala wyrażać utratę informacji jako różnię wartości *LR* uzyskanej przez biegłego z wartością *LR* obliczoną na podstawie wiedzy o prawdziwym stanie rzeczy. Dzięki tej procedurze ocenia się nie tyko właściwości dyskryminujące wyliczanych wartości *LR*, ale również wskazuje, czy proces kalibracji stosowano czy też nie i jednocześnie określa stopień utraty informacji. Przykład oceny metodyki obliczania wartości *LR* w systemie rozpoznawania mówcy zaprezentowano na przykładzie zastosowania przejrzystego i zestandaryzowanego protokołu postępowania w przypadku rozpoznawania mówcy (NIST 2006 SRE), co zilustrowano w formie wykresów DET. Podsumowując, proponowany przykład procedury oceny jakości uzyskiwanych wartości *LR* cechuje się przejrzystością i możliwością kontroli.

Podziękowania
Badania były finansowane przez Hiszpańskie Ministerstwo Nauki i Technologii w ramach projektu TEC2006-13170-C02-01. Autorzy pragną podziękować prof. Colinowi Aitkenowi (Uniwersytet Edynburski, Wielka Brytania) i dr. Grzegorzowi Zadorze (Instytut Ekspertyz Sądowych, Kraków) za pomocne komentarze i dyskusję.