ELSEVIER

Speech Communication 31 (2000) 193–203

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

The inference of identity in forensic speaker recognition

Christophe Champod *, Didier Meuwly

Institut de Police Scientifique et de Criminologie, University of Lausanne, CH-1015 Lausanne, Switzerland

Received 28 October 1998; received in revised form 13 September 1999

Abstract

The aim of this paper is to investigate the ways of interpreting evidence within the field of speaker recognition. Several methods – speaker verification, speaker identification and type I and type II errors statement – will be presented and evaluated in the light of judicial needs. It will be shown that these methods for interpreting evidence unfortunately force the scientist to adopt a role and to formulate answers that are outside his scientific province. A Bayesian interpretation framework (based on the likelihood ratio) will be proposed. It represents an adequate solution for the interpretation of the aforementioned evidence in the judicial process. It fills in the majority of the gaps of the other inference frameworks and allows likening the speaker recognition to the same logic than the other forensic identification evidences. © 2000 Published by Elsevier Science B.V. All rights reserved.

Zusammenfassung

Das Ziel dieses Artikel ist die Schlussfolgerungsschemen zu untersuchen, welche im Bereich der Sprechererkennung verwendet werden. Mehrere Methoden (Überprüfung des Sprechers, Identifikation des Sprechers, Darlegen der Fehler vom Typ I und II) werden in bezug auf die gerichtlichen Bedürfnisse vorgestellt und bewertet. Diese Analyse zeigt auf, dass diese Interpretationslösungen den Experten oft zwingen eine Rolle anzunehmen und Fragen zu beantworten, welche leider die reinen wissenschaftlichen Kompetenzen übersteigen. Ein Interpretationsrahmen vom Typ Bayes (basierend auf dem Wahrscheinlichkeitsverhältnis - likelihood ratio) wird präsentiert und als eine angemessene Lösung zur Interpretation des Beweises im Gerichtsprozess vorgeschlagen. Dieses Vorgehen füllt die Mehrheit der Logiklücken der anderen Schlussfolgerungssysteme auf und erlaubt die Sprechererkennung in das gleiche Interpretationsschema wie die anderen forensischen Identifikationsbeweise einzuordnen. © 2000 Published by Elsevier Science B.V. All rights reserved.

Résumé

Le but de cet article est d'investiguer les schémas d'inférence en place dans le domaine de l'identification du locuteur. Plusieurs canevas d'interprétation (vérification de locuteur, identification de locuteur, présentation des erreurs de premier et de second type) vont être présentés et évalués en regard des besoins judiciaires. Il ressortira de cette analyse que ces solutions proposées pour l'interprétation force souvent l'expert à adopter un rôle et à répondre à des questions qui vont malheureusement au-delà des seules données scientifiques. Un cadre d'interprétation de type Bayesien (basé sur le rapport de vraisemblance) sera présenté et proposé comme une solution adéquate pour l'interprétation de la preuve

* Corresponding author. Present address: The Forensic Science Service, Metropolitan Laboratory, 109 Lambeth Road, London SE1 7LP, UK. Tel.: +0171 230 6732; fax: +0171 230 6432.

E-mail addresses: christophe.champod@ipsc.unil.ch (C. Champod), didier.meuwly@ipsc.unil.ch (D. Meuwly).

PII: S0167-6393(99)00078-3

dans le procès. Ce canevas comble la majorité des lacunes logiques des autres systèmes d'inférence et permet d'inscrire l'identification de locuteur dans le même schéma d'inférence que les autres moyens d'identification forensiques.

Keywords: Forensic speaker recognition; Inference; Bayesian approach

1. Introduction

The aim of forensic speaker recognition is to help the police to link recordings made in connection with criminal activities (anonymous calls or telephone tapping) to one or more suspects. Different methods can be applied either singly or combined (perceptual, visual or computer study (Majewski and Basztura, 1996) or even linguistic study) to determine if the unknown voice belongs to the suspect, however our goal is not to discuss the respective value and validation of these methods.

The method(s) will provide a probabilistic statement, objective or subjective, which attempts to give the court an indication of the strength of the evidence. In this paper, we would like to study the process of identification, which is completely independent of the method chosen.

When evidence associates an unknown voice to a suspect, the following question – common to most identification evidence – is: *What is the probability that this evidence* (*voice*) *came from that person*? Some experts do not hesitate answering it but, most of the time, they do not have a clear understanding of the inferential process and the respective duties of the actors involved in the judicial process, jurists, experts, members of the jury, etc.

The debate on the interpretation of identification evidence has been initiated by the famous case *People v. Collins* (Fairley and Mosteller, 1974) and it remains intense among forensic practitioners (see for example (Kind, 1994) and a very curt response (Robertson and Vignaux, 1994a)). It would be wrong to argue that a compromise has been achieved, but the reflections may guide or promote discussion in the field of speaker recognition.

We propose to discuss forensic speaker recognition in the light of current literature related to the interpretation of forensic evidence. The concept of identity used in criminalistics will be defined, then proposed guidelines for making decisions applied in the field of speaker recognition will be presented and discussed in a forensic context. This discussion will profit from the abundant literature related to DNA and trace evidence. Finally, an interpretation framework, which relies on Bayes' theorem, will be outlined as it becomes more and more accepted in other fields of forensic science including speaker recognition (Broeders, 1995; Lewis, 1984).

2. The concept of identity in criminalistics

In criminalistics, the identification process seeks individualisation (Tuthill, 1994). Identifying a person or an object means that it is possible to distinguish this person or object from all others on the surface of the earth. The forensic individualisation process can be seen as a reduction process beginning from an initial population to a single person. In fields like fingerprints, shoeprints, toolmarks and firearms, the size of the initial population is set to its maximum, the population on earth. The reduction factor comes from the specificity or rarity of the concordant features observed between the trace (i.e., a papillary mark recovered on the crime scene) and the control material (i.e., an inked fingerprint from the suspect). The conclusion of an identification is an opinion, a statement of probability, objective and/or subjective, expressing that the chance of observing on earth another person or object presenting the same characteristics is nil.

The attention of jurists or scientists is too easily concentrated upon the ability of a technique to end up with an absolute certainty. If the technique can positively conclude an identification, it is greeted as a panacea, if it cannot, it is damned as unreliable. This ignores the vital point that any technique will only function to a high degree of precision under controlled conditions and the conditions under which forensic scientists work are far from ideal. It follows that, in many cases, a forensic scientist will not be able to provide a definitive answer but only a probabilistic figure or opinion (Robertson and Vignaux, 1995). If an ultimate set of specific features is not present or not detected in the evidence, then the criminalist will not provide an identification but will express a probability statement, verbally or numerically, which attempts to assess the value of the evidence. Indeed the identification process remains in essence a statistical process based on objective data and/or subjective assessment related to the expert's experience.

In some forensic fields (fingerprints (Anon, 1980), toolmarks (AFTE Committee, 1992), shoeprints (Bodziak, 1990)), practitioners have voluntarily excluded probability statement – other than exclusion and identification – from their conclusions. All pieces of evidence between these extremes are classified as inconclusive. We sustain that there is no logical reason to suppress probability statements and believe that the refusal of qualified opinions is a policy decision, even if the distinction of the arguments (policy or scientific argumentation) is not so clear in the literature (Robertson and Vignaux, 1994b) (appreciate the dogmatic statement proposed recently by an FBI working group on fingerprint evidence: "Friction ridge identifications are absolute conclusions. Probable, possible or likely identification are outside the acceptable limits of the science of friction ridge identification") (TWGFAST, 1997, p. 432).

We agree with Lempert (1977) in the sense that each piece of evidence is relevant if it tends to make the matter which requires proof more or less probable than otherwise. Hence, a piece of evidence which only approaches the absolute identification constitutes relevant evidence which should not be ignored.

3. Definitions

Let us consider a hypothetical vocal recording (X) and a control recording (S) of a suspect.

The comparison of the unknown utterance X and the suspect's sample S leads the scientist either to a numerical assessment which describes the distance between them, or to a subjective opinion stating their similarities and differences. This represents the evidence and will be generally noted as E. E may be a numerical value expressing a random match probability or a subjective opinion on the frequency of the set of attributes. The ultimate question relies on the evaluation of the probative value of this evidence.

As no standardised framework to interpret such evidence has been accepted in the field of forensic speaker recognition, we question if the task of the expert should be speaker verification (discrimination task), speaker identification (classification task) (O'Shaughnessy, 1986) or if the scientist should only report type I and type II errors.

4. Proposed frameworks used to assess the evidence

We believe that the underlying logical framework should be the same for both kinds of evidence, objective or subjective and that the same rules should govern the interpretation of statements like "the random match probability is 1 in a million" or "the frequency of these features is very very low".

4.1. Speaker verification

Speaker verification (discrimination task) has been proposed by Doddington (1985) for forensic speaker recognition. The decision of discrimination between the unknown recording X and the control recording S depends on a threshold, that can be qualitative (subjective assessment of similarities and differences) or quantitative (closeness numerically expressed): discrimination is interpreted as an exclusion and non-discrimination as an identification.

The above concept of identity does not correspond to the definition of forensic individualization; if the random match probability is not nil (corollary of the threshold), the conclusion of "the suspect S is identified" is inadequate and misleading.

Moreover, it must be pointed out that the threshold is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert. But jurists will interpret this threshold as an expression of the criminal standard "beyond reasonable doubt". *Would jurists accept that the concept of reasonable doubt on the identification of a suspect escapes their province and that the threshold is imposed onto the court by the scientist?* The response in the doctrine is negative, as expressed by the members of the panel on statistical assessments as evidence in courts: "[...] the law may establish different thresholds for what is sufficient evidence in a case from those that statisticians would normally require in drawing conclusions. Clearly, the law must prevail and the statistician must adjust to the law's standards. Put another way, it is the utility function of the court that is appropriate, not the utility function of the statistician" (Fienberg, 1989, p. 141).

Therefore, speaker verification is clearly inadequate for forensic purposes, because it forces the scientist to adopt a role and to make decisions which are devolved upon the court.

4.2. Speaker identification

Speaker identification (classification task) has been proposed in an open-set by Künzel (1994) for forensic purposes. In fact, the classification cannot take place in a closed-set of speakers (closed-set identification) because the assessment of the credibility of the exhaustiveness of the number of suspects is outside the duties of the expert; it is a judicial matter pertaining to the court. In addition, it seems particularly unfair to disclose only the identity of the best candidate without providing the evidence obtained for the others. To illustrate this concern, let us consider a case of burglary in which a window has been broken by the perpetrator; two suspects are apprehended and tiny glass fragments are recovered on their respective garments. A measure of the refractive index of the glass shows some concordance with the broken window with the first suspect having a random match probability of 1/1100 and with the second having a random match probability of 1/900 (we have adopted a continuous approach according to (Walsh et al., 1996)). If, following the closed-set identification task, only one of the two is declared identified, it focuses the evidence in a very misleading way on the first suspect. Indeed the evidence does not enable the strong favourisation of one or the other hypotheses that may be more relevant to the court: H_1 the first suspect is the offender as opposed to H_2 the second suspect is the offender.

To overcome this default, the classification should then take place in an open-set of speakers (open-set identification), but such a framework still implies a final discrimination decision based on a threshold and suffers from the same conceptual drawbacks as the verification task.

4.3. Reporting type I and type II errors

Statisticians often make judgements based on the costs of the two types of errors (type I and type II), using methods like receiver operating characteristic (ROC) to measure the performance in decision tasks involving physical instruments and human observers (Committee on the Evaluation of Sound Spectrograms, 1979).

If the choice is focused on speaker verification or open-set speaker identification, the decision task can be represented by a table (Table 1) with the possible states and possible decisions.

Such a performance measure has been proposed in forensic speaker recognition to assess automatic (Paoloni et al., 1994) or voicegram examiners (Committee on the Evaluation of Sound Spectrograms, 1979).

Let us consider the following example: the utterance of a suspect is compared to the unknown recording and accepted by a system with the performance described in Table 1. On learning that the result is positive, the expert may well draw the following conclusion: the suspect is very probably

Table 1
Possible states and decisions

	State	
Decision	Correct	Wrong
Positive (+)	0.99	0.01 (type II)
Negative (−)	0.01 (type I)	0.99

the author (an analogous discussion is given by (Royall, 1997, pp. 1–5)). *Is such a conclusion valid?*

Let us restate this conclusion in terms of probability: the suspect is very probably identified (ID), given the positive decision (response given by the system), that means that $\Pr(\mathrm{ID}\,|\,+)$ is around 0.99. According to Bayes' theorem, the latter probability depends partly on the positive result but also on the prior (before the analysis) probability of the fact (the identification), Pr(ID). In fact, we can write

$$\begin{aligned}\Pr(\mathrm{ID} \mid +) &= \frac{\Pr(+ \mid \mathrm{ID})\ \Pr(\mathrm{ID})}{\Pr(+ \mid \mathrm{ID})\ \Pr(\mathrm{ID}) + \Pr(+ \mid \bar{\mathrm{ID}})\ \Pr(\bar{\mathrm{ID}})} \\ &= \frac{0.99\ \Pr(\mathrm{ID})}{0.99\ \Pr(\mathrm{ID}) + 0.01(1-\Pr(\mathrm{ID}))}. \qquad (1)\end{aligned}$$

The validity of the conclusion depends critically on the prior probability on the identification. A similar demonstration has been made for the interpretation of DNA evidence in forensic science (Balding and Donnelly, 1994). As shown in Fig. 1 (Berry, 1991), the statement that $\Pr(\mathrm{ID}\,|\,+) > 0.99$ is correct only if $\Pr(\mathrm{ID}) \geqslant 0.5$.

Therefore, it would be wrong to state that the identity of the speaker is demonstrated with an error rate of 1%, because this conclusion only holds if the expert assumes (generally without the adequate knowledge or even awareness) a prior probability of identity of 1/2! It could be claimed that this practice is "neutral" because if the alleged speaker is not the true speaker, only one other person has to be considered; see a similar argument in paternity cases in (Hummel, 1984). This policy is arbitrary and misleading at best and any other value of prior probability is equally misleading and usurps the role of the judge and of the jury (Taroni and Aitken, 1996b). Even if it is very tempting when assessing evidence to try to determine a value for the probability of identification, this is the role of the jury and/or judge and not the role of the forensic scientist (Evett, 1983). A binary decision on the identification provided by the expert is thus inadequate for forensic speaker recognition.

To overcome this impossibility, many scientists will argue that the adequate way of presenting the results would be simply to state the type II error and explaining to the court that it represents the chance of obtaining a positive result whereas the suspect is not involved – percentage of false positives $\Pr(+ \mid \overline{\mathrm{ID}})$ – and not the probability of a false identification given a positive decision $\Pr(\overline{\mathrm{ID}} \mid +)$. Italian courts have already accepted evidence presented using a type II error:

> [...] un nuovo sistema d'analisi definito IDEM e di nuove metodologie statistiche e che presenterebbe un "potere risolutivo", cioè una capacità di identificazione della voce, fino a 1:1 000 000 (Sez. V, 1994). [1]

As type II error is only a measure of average probability, forensic scientists would propose not to use them blindly in all cases and, when possible, to give the court an estimate of a random match probability specific to the case (Stoney, 1991).

This policy will unfortunately not take into account the fact that psychological research has underlined the fallacious way in which people

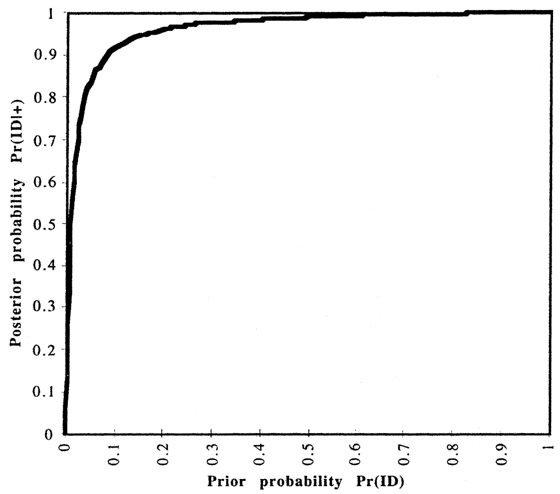

Fig. 1. Evolution of the posterior probability in function of the prior probability.

[1] Free translation: "a new analysis system called IDEM and new statistical methodologies reach a "discrimination power" of 1:1 000 000 for forensic speaker identification".

reason when dealing with uncertainty and probabilities, especially with conditional probabilities (Fienberg and Finkelstein, 1996). Intuition is abad substitute for the laws of probability in evaluating uncertainty (Evett, 1993). Thompson and Schumann (1987) have exposed two kinds of dangers linked to a misinterpretation of probabilities at trial: the prosecutor's and the defence's fallacy. Several papers on this specific subject were published in judicial and scientific reviews showing other subtle forms of fallacy (Kaye, 1993; Koehler, 1993; Evett, 1995). Victims of this confusion are both jurists and experts (Champod and Taroni, 1993, 1994; Taroni and Aitken, 1997)!

Imagine a case where a juror hears that the probability of an innocent person possessing the features of the unknown recording is 1 in a 100. What can be deduced correctly from this value?

Here we enter into a well-known area of forensic science, which is most exemplified by the controversies around the interpretation of statistical data: DNA evidence (Koehler, 1993). The value above can be interpreted from two points of view: that of the prosecution and that of the defence:

1. The juror or even the scientist in the witness box, may misconstrue (sometimes with the help of the prosecution) this figure to mean that 99% (1 – 1/100) represents the probability that the suspect is implicated.
2. On the other hand, the defence can argue that in a city of 100 000 people, the statistical figure of 1 in a 100 will, on average, leave us with 1000 persons, amongst them the defendant. The probability of a correct identification of the defendant would then be 1/1000 and not 99%.

These two dramatically divergent interpretations of the same probability value force us and the court to comprehensively understand the inferential process involved:

Argument (1) constitutes an elementary but common error called the "prosecutor's fallacy" (Thompson and Schumann, 1987), the inversion fallacy (Kaye, 1993) or more generally the error of transposing the conditional (Evett, 1995). This fallacious argument leads the jury to interpret the probability without any knowledge of the prior probability of implication. In mathematical terms, the fallacy is

$$\Pr(\mathrm{ID} \mid E) = 1 - \Pr\left(E \mid \overline{\mathrm{ID}}\right). \tag{2}$$

It consists of replacing $\Pr(\overline{\mathrm{ID}} \mid E)$ by $\Pr(E \mid \overline{\mathrm{ID}})$, therefore transposing the conditional. The correct equation is therefore

$$\Pr(\mathrm{ID} \mid E) = 1 - \Pr\left(\overline{\mathrm{ID}} \mid E\right). \tag{3}$$

As it has been correctly pointed out, even by courts of law (see the case R. v. Deen quoted by Redmayne (1995)), there are two distinct questions under scrutiny:

(a) What is the probability that an individual would match the unknown recording given he is innocent?
(b) What is the probability that an individual is innocent, given that he matches the unknown recording?

The prosecutor's fallacy consists of giving the answer to the first question (a) as the answer to the second (b).

The defence argument (2) is known as the defence attorney's fallacy (Thompson and Schumann, 1987). It attempts to make the court believe that a large number of possible suspects can be at the origin of the unknown recording. This argument is certainly misleading, because the number of possible authors considered by the defence is generally with no relation to the reality of the case, yet it does not incorporate any logical error in the inference.

The adequate interpretation of the value of the evidence provided needs to consider the statistical value obtained in a different framework, namely, the Bayesian framework, which conversely helps forensic scientists, jurists and members of the jury in reaching their conclusions.

5. Bayesian framework for the evaluation of evidence

The approaches we have referred to up to now may be qualified as prescriptive approaches according to Rudram (1996) drawing conclusions on the facts. We will compare them to a Bayesian approach – or likelihood approach – which leads to statement of the degree of support for one hypothesis versus another (see the response to Rudram by Taroni and Aitken (1996a)). As pointed out by Lewis (1984), who in 1984 proposed the use of Bayes theorem in speaker identification, evidence does not consist uniquely of scientific data. In general, science can only provide additional information to assist an answer that must be ultimately arrived at inductively. The forensic individualisation process is best explained at present by the hypothetical-deductive method (Kwan, 1977).

We believe that a probabilistic model – Bayes' theorem – is a useful tool

1. for assisting scientists to assess the value of scientific evidence;
2. to help jurists to interpret scientific evidence;
3. to clarify the respective roles of scientists and of members of the court.

The Bayesian model allows the revision based on new information (E) of a measure of uncertainty about the truth or otherwise of an issue. This approach is especially useful with scientific evidence (Aitken and Stoney, 1991; Aitken, 1995; Robertson and Vignaux, 1995). It shows how data can be combined with prior background knowledge (I) and new data in order to give posterior probabilities for particular outcomes or issues. We simplify the issue to the comparison of two exclusive hypotheses:

ID: the unknown recording X was made by the suspect;

$\overline{\mathrm{ID}}$: the unknown recording was not made by the suspect.

However, these issues are not always as straightforward or exhaustive as could be deduced from our notation.

The court is interested in the explanations given for the evidence. We could consider systematically providing two mutually exclusive and exhaustive hypotheses, such ID and $\overline{\mathrm{ID}}$, as done in previous examples. *But who should provide the possible explanations*? We have often felt that forensic scientists or laboratory managers were trying to define general issues for every case like the hypotheses ID and $\overline{\mathrm{ID}}$. Referring to Bayes theorem, we observe that the issues are set before the evaluation of evidence. In fact the estimation of prior probabilities requires the knowledge of the hypotheses involved. Consequently, the definition of the hypotheses themselves are outside the duties of the experts: it is thus a matter of court.

The prosecution will normally present the evidence as a result of a criminal activity of a suspect. This hypothesis will be noted H_1. Nevertheless, this event is rarely the only possible explanation of the evidence and the forensic scientist must also consider the explanation(s) that will be provided by the defence: hypotheses noted $\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3, \ldots, \mathrm{H}_N$ (Robertson and Vignaux, 1995). At a particular moment in a trial, the context is restricted to two competitive hypotheses: that proposed by the prosecution and that of the defence. Therefore, the context of the interpretation of the evidence is not defined by the forensic scientist but by the prosecution and defence as a function of the specific circumstances of the case. As an example, the following set of explanations could be put to the court by the parties:

H_1: the suspect is the author of the unknown recording.

H_2: the speaker at the origin of the unknown recording is not the suspect but an unknown man, according to the fundamental frequency of the unknown recording.

H_3: the speaker at the origin of the unknown utterance is not the suspect but an unknown woman described by a reliable witness.

The evidential value of the forensic examination consists of the assessment of the probabilities of the observations under two competitive hypotheses. This could be H_1 against H_2, H_1 against H_3 or whichever other competitive hypothesis is expressed by the defence. This means that the interpretation of the evidence will change as a function of the scenarios proposed by the opposing parties. The hypotheses are defined in the light of background information I which is derived from police inquiries (witness testimonies, criminal history records, etc.) and represents data other than the evidence E itself. This information can greatly modify the interpretation of the evidence. Consequently, we advocate a complete co-operation between the scientist, the case agent and the court to be aware of the alleged circumstances of the case.

5.1. What judicial question does the court require to be solved in collaboration with the forensic scientist?

In general, the court wants to know the odds that this suspect has produced the recording X given the circumstances of the case (I) and the observations made by the forensic scientist (E). In mathematical terms, the court looks for the odds $\mathrm{O}(\mathrm{H}_1 \mid E, I)$ on an issue H_1 versus its alternative H_2. These odds are posterior odds.

$$\mathrm{O}(\mathrm{H}_1 \mid E, I) = \frac{\Pr(\mathrm{H}_1 \mid E, I)}{\Pr(\mathrm{H}_2 \mid E, I)}. \tag{4}$$

5.2. What is the information given by the scientist?

The statement generally expresses the probability of the evidence if the suspect has not produced the unknown recording $\Pr(E \mid \mathrm{H}_2, I)$.

The Bayesian formula (5) helps to relate these two questions, as it shows how the prior odds are modified by the evidence to obtain the posterior odds (the judicial question):

$$\begin{aligned}\mathrm{O}(\mathrm{H}_1 \mid E, I) &= \frac{\Pr(\mathrm{H}_1 \mid E, I)}{\Pr(\mathrm{H}_2 \mid E, I)} \\ &= \frac{\Pr(E \mid I, \mathrm{H}_1)}{\Pr(E \mid I, \mathrm{H}_2)} \cdot \frac{\Pr(\mathrm{H}_1 \mid I)}{\Pr(\mathrm{H}_2 \mid I)},\end{aligned}$$

$$\underbrace{\mathrm{O}(\mathrm{H}_1 \mid E, I)}_{\text{posterior odds}} = \underbrace{\frac{\Pr(E \mid I, \mathrm{H}_1)}{\Pr(E \mid I, \mathrm{H}_2)}}_{\text{likelihood ratio}} \cdot \underbrace{\mathrm{O}(\mathrm{H}_1 \mid I)}_{\text{prior odds}}. \tag{5}$$

The prior odds are an evaluation of the weight attributed to H_1 versus its alternative H_2 before the forensic examination. Here, the fact finder will take into account all background information that has been collected before the forensic expertise. For example, data from police investigations, eyewitness statements or data from the criminal history record of the suspect will contribute to background information.

The likelihood ratio (LR) measures the value of the evidence, it summarizes the expert's statement as a ratio of probabilities:

$$\mathrm{LR} = \frac{\Pr(E \mid I, \mathrm{H}_1)}{\Pr(E \mid I, \mathrm{H}_2)} \tag{6}$$

(or in some cases probability densities).

The numerator of (6) is the probability of the evidence given that this suspect is indeed at the origin of the recording X. Generally, this value calls for an assessment of the intra-variability of the system. Ideally, it would approach a value close to 1. The denominator of (6) is the probability of the evidence given that the unknown recording was in fact not produced by the suspect. This value is the random match probability. It can be derived from an objective or subjective estimation of the relative frequency of the concordant features in the relevant population. The adequate population to be surveyed will be defined by a population corresponding to the possible authors of the unknown recording (Lempert, 1993). The definition of H_2 excludes the implication of the defendant, hence whatever specificity the defendant's voice may have is irrelevant. This point helps to argue against irrelevant defence argument in the form: ''you have provided a frequency estimate based on the Swiss population, whereas the suspect is not a Swiss native but comes from south of France, consequently the evidence provided is not applicable in this case.'' (see the analogous discussion for DNA evidence by (Evett and Weir (1991)).

Alone, the LR is not sufficient to state on the issue, it expresses how much more likely the evidence is under one hypothesis versus the other. The calculation of the LR, however, is not a ''Bayesian analysis'', as this term usually implies the assignment of prior probabilities.

The scientist is generally not in a position to assess the odds in favour of an issue, because a complete assessment must combine both the forensic statement (E) and background information (I). The scientist does not usually have access to the background information that is available to a member of a jury or a judge. Most of the time the scientist does not know the other pieces of evidence in a case and thus is not able to correctly assess the prior odds. This means that the numerical statement or the opinion alone given by the scientist is not sufficient to assess the final odds on H_1. Scientists are concerned solely with the likelihood ratio. It is jurists who deal with the odds of the issue. As shown in the other frameworks, this distinction of roles is generally not realised. Hence, statements of scientific certainty on the issue itself $\Pr(H_1 \mid E, I) = 1$ or even probabilistic statement on the issue $\Pr(H_1 \mid E, I) = p$, offer the court the wrong answer (the opinion of the forensic scientist on an issue) to the right question (the question of the court) (Taroni and Aitken, 1996a). Some examples of such usurped statements can be found in the literature:

> "From the evidence collected – which was presented and discussed here – it was possible to safely conclude that the voice present in the recording under investigation was the same one in the former minister's interviews and the tape has not been edited." (Figueiredo et al., 1995, p. 43.)

Forensic scientists should give the court an evaluation which illustrates the convincing force of the results (Kaye, 1992). This evaluation is made through an assessment of a likelihood ratio which is applied to pairs of hypotheses, indicating when a given set of observations is evidence for one versus the other. This need for restricting experts' conclusions to likelihood ratios has already been identified in various fields of forensic science. In 1904, Darboux, Appell and Poincaré wrote in the famous French criminal case of Dreyfus (Taroni et al., 1998):

> "Since it is absolutely impossible for us to know the a priori probability, we cannot say: this coincidence proves that the ratio of the forgery's probability to the inverse probability is a real value. We can only say that, following the observation of this coincidence, this ratio becomes X times greater than before the observation."

This way of expressing a conclusion, giving only the likelihood ratio, is – 90 years later – still presented as the only one authorised for forensic scientists (we refer to (Robertson and Vignaux, 1992) for paternity cases, (Huber, 1980) and (Hilton, 1995) for document examinations, (Dienet, 1984) for toolmarks, (Stoney, 1985) for fingerprints more generally, (Aitken, 1995) for trace evidence and (Robertson and Vignaux, 1995) for forensic evidence in general).

Hence, the concept of evidence is essentially relative. It explains how observations should be interpreted as evidence for H_1 vis-à-vis H_2, but it makes no mention of how those observations should be interpreted as evidence in relation to H_1 alone (Royall, 1997).

Using this Bayesian framework, the two main fallacies described earlier can be avoided:

1. The prosecutor's fallacy is avoided since $\Pr(E \mid H_2, I)$ influences only the likelihood ratio and not the prior odds. To obtain the posterior odds obtained following the fallacious argument, we would have to impose prior odds of 1 against 1, as if only two persons were with equal probability involved in the case.
2. The defence's fallacy on the other hand, asks to adopt prior odds where the suspect is considered as likely to be involved as a relevant population which is excessively enlarged. Again, most of the time, this is not a realistic description of the case.

6. Conclusion

We have shown the inadequacy of the main solutions proposed to assess the evidence in the field of forensic speaker recognition. The concept of identity underlying the verification and the identification tasks (closed-set and open-set) does not correspond to the concept of identity accepted in forensic science. In addition, the use of these guidelines force the scientist to deal – without being aware – with prior and/or posterior probabilities on the issue of identification itself, whereas these are assessments pertaining only to the court. On the other hand, reporting only type II errors (or random match probabilities) can lead to fallacious interpretation either from jurists or scientists.

The Bayesian interpretation framework overcomes most of these difficulties and provides a coherent way of assessing and presenting this kind of evidence. It provides the forensic scientist with guidelines for the evaluation of scientific evidence. These can be summarised by the following points:

- The scientific value of forensic evidence is best assessed with a likelihood ratio, that forces the scientist to consider the evidential value under two competitive hypotheses provided by the court and not on the value of the issue itself.
- The numerator of the likelihood ratio $\Pr(E \mid H_1, I)$ requires the probability of the evidence that the suspect is indeed the author. This probability is not automatically equal to one and must be assessed in each case taking into account the intra-variability of the whole process (collection, analysis and result) which provides the evidence.
- The denominator of the likelihood ratio $\Pr(E \mid H_2, I)$ is best estimated through a relative frequency of the concordant features in the relevant population. The relevant population is dictated by the hypothesis proposed by the defence and by background information related to the case.

In conclusion, the concept of evidence is essentially relative to the case and its value is best expressed using a likelihood ratio. Referring to the definition of this likelihood ratio, the analysis of the scientific evidence does not allow the scientist alone to make an inference on the identity of the speaker.

References

AFTE Committee, 1992. Theory of identification range of striæ comparison reports and modified glossary definition an AFTE criteria for identification committee report. AFTE Journal 24, 336.

Aitken, C.G.G., 1995. Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, Chichester, UK.

Aitken, C.G.G., Stoney, D.A., 1991. The Use of Statistics in Forensic Science. Ellis Horwood, Chichester, UK.

Anon, 1980. Resolution VII amended. Identification News 30, 3.

Balding, D.J., Donnelly, P., 1994. The prosecutor's fallacy and DNA evidence. Criminal Law Review, 711–721.

Berry, D.A., 1991. Probability of Paternity. In: Aitken, C.G.G., Stoney, D.A. (Eds.), The Use of Statistics in Forensic Science. Ellis Horwood, New York, pp. 150–156.

Bodziak, W.J., 1990. Footwear Impression Evidence. Elsevier, New York.

Broeders, A.P.A., 1995. The role of automatic speaker recognition techniques in forensic investigations. In: Proceedings of the International Conference of Phonetic Science. Vol. 3, pp. 154–161.

Champod, C., Taroni, F., 1993. Les préjugés de l'Accusation ou de la Défense dans l'évaluation de la preuve technique. Revue pénale suisse 111 (3), 223–235.

Champod, C., Taroni, F., 1994. Probabilités au procès pénal – risques et solutions. Revue pénale suisse 112 (2), 194–219.

Committee on the Evaluation of Sound Spectrograms, N.R.C., 1979. On the Theory and Practice of Voice Identification. National technical Information Service, Washington DC.

Dienet, W., 1984 The applications of probabilistic methods in forensic science examination. In: The 10th IAFS triennial meeting, Oxford, UK.

Doddington, G.R., 1985. Speaker recognition – Identifying people by their voices. Proceedings of IEEE 73 (11), 1651–1664.

Evett, I.W., 1983. What is the probability that this blood came from that person? A meaningful question. Journal of the Forensic Science Society 23, 35–39.

Evett, I.W., 1993. Interpretation of evidence. International Criminal Police Review 444, 12–17.

Evett, I.W., 1995. Avoiding the transposed conditional. Science and Justice 35 (2), 127–131.

Evett, I.W., Weir, B.S., 1991. Flawed reasoning in court. Chance: New Directions for Statistics and Computing 4 (4), 19–21.

Fairley, W.B., Mosteller, F., 1974. A conversation about Collins. The University of Chicago Law Review 41, 242–253.

Fienberg, S.E., 1989. The Evolving Role of Statistical Assessments as Evidence in the Courts. Springer, New York. Vol. 1, p. 357.

Fienberg, S.E., Finkelstein, M.O., 1996. Bayesian Statistics and the Law. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics. Oxford University Press, Oxford, Vol. 5, pp. 129–146.

Figueiredo, R.M., Palhares, F.A.B., Nagle, E.J., 1995. Description of a real case involving the identification of speaker and the authentification of a magnetic tape recording. In: Jacob, B., Bonte, W. (Eds.), Advances in Forensic Sciences. Verlag Dr. Köster, Berlin. Vol. 3, pp. 35–44.

Hilton, O., 1995. The relationship of mathematical probability to the handwriting identification problem. International Journal of Forensic Document Examiners 1 (3), 224–229.

Huber, R.A., 1980. The quandary of "qualified" opinions. Canadian Society of Forensic Science Journal 13 (3), 1–8.

Hummel, K., 1984. On the theory and practice of Essen–Möller's W value and Gürtler's paternity index. Forensic Science International 25, 1–17.

Kaye, D.H., 1992. Proof in law and science. Jurimetrics Journal 32 (3), 313–321.

Kaye, D.H., 1993. DNA evidence: Probability, population genetics, and the courts. Harvard Journal of Law and Technology 7 (1), 101–172.

Kind, S.S., 1994. Crime investigation and the criminal trial: A three chapter paradigm of evidence. Journal of the Forensic Science Society 34, 155–164.

Koehler, J.J., 1993. Error and exaggeration in the presentation of DNA evidence at trial. Jurimetrics Journal 34 (1), 21–39.

Künzel, H., 1994. Current approaches to forensic speaker recognition. In: Proceedings of the ESCA Workshop on Automatic Speaker Recognition Identification Verification, Martigny, Switzerland.

Kwan, Q.Y., 1977. Inference of identity of source. Ph.D. thesis of Criminology. Department of Forensic Science, University of California, Berkeley, CA, USA.

Lempert, R.O., 1977. Modeling relevance. Michigan Law Review 75, 1021–1057.

Lempert, R.O., 1993. The suspect population and DNA identification. Jurimetrics Journal 34 (1), 1–7.

Lewis, S.R., 1984. Philosophy of speaker identification. Police applications of speech and tape recording analysis. Proceeding of the Institute of Acoustics 6 (1), 69–77.

Majewski, W., Basztura, C., 1996. Integrated approach to speaker recognition in forensic applications. Forensic Linguistics 3 (1), 50–64.

O'Shaughnessy, D., 1986. Speaker recognition. IEEE ASSP Magazine 3, 4–17.

Paoloni, A., Falcone, M., Bimbot, F., Chollet, G., 1994. Outline of a comprehensive assessment methodology for speaker recognition task in forensic application. In: Annual Meeting of the International Association For Forensic Phonetics. Cardiff, 9–13 July.

Redmayne, M., 1995. Doubts and burdens: DNA evidence, probability and the courts. Criminal Law Review, 464–482.

Robertson, B., Vignaux, G.A., 1992. Unhelpful evidence in paternity cases. New Zealand Law Journal 9, 315–317.

Robertson, B., Vignaux, G.A., 1994a. Crime investigation and the criminal trial. Journal of the Forensic Science Society 34 (4), 270.

Robertson, B., Vignaux, G.A., 1994b. The interpretation of fingerprints. Expert Evidence 3 (1), 3–8.

Robertson, B., Vignaux, G.A., 1995. Interpreting Evidence – Evaluating Forensic Science in the Courtroom. Wiley, Chichester, UK.

Royall, R.M., 1997. Statistical Evidence – A Likelihood Paradigm. Chapman & Hall, London.

Rudram, D.A., 1996. Interpretation of scientific evidence. Science and Justice 36 (3), 133–138.

Sez. V, l., 1994. Massimario No. 23. La Giustizia Penale 35 (1 (Gennaio)), 42.

Stoney, D.A., 1985. Quantitative assessment of fingerprint individuality. Ph.D. dissertation. Graduate Division, University of California, Berkeley, CA.

Stoney, D.A., 1991. Transfer evidence. In: Aitken, C.G.G., Stoney, D.A. (Eds.), The Use of Statistics in Forensic Science. Ellis Horwood, New York, pp. 27–50.

Taroni, F., Aitken, C.G.G., 1996a. Interpretation of scientific evidence. Science and Justice 36 (4), 290–292.

Taroni, F., Aitken, C.G.G. 1996b. Probabilities and DNA Evidence in Paternity and Criminal Cases. Judicial Questions and Fallacious Answers. 75. Jahrestagung der Deutschen Gesellschaft für Rechtsmedizin, Zürich. 24–28 September 1996.

Taroni, F., Aitken, C.G.G., 1997. Forensic science at trial. Jurimetrics Journal 37, 327–337.

Taroni, F., Champod, C., Margot, P.A., 1998. Forerunners of Bayesianism in early forensic science. Jurimetrics Journal 38, 183–200.

Thompson, W.C., Schumann, E.L., 1987. Interpretation of statistical evidence in criminal trials the prosecutor's fallacy and the defense attorney's fallacy. Law and Human Behavior 11 (3), 167–187.

Tuthill, H., 1994. Individualization: Principles and Procedures in Criminalistics. Lightning Powder, Salem.

TWGFAST, 1997. Technical working group on friction ridge analysis, study and technology (TWGFAST) proposed guidelines. Journal of Forensic Identification 47 (4), 423–437.

Walsh, K.A.J., Buckleton, J.S., Triggs, C.M., 1996. A practical example on the interpretation of glass evidence. Science and Justice 36 (4), 213–218.