Science and Justice 56 (2016) 42–57

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case

Ewald Enzinger [a,b,c,*], Geoffrey Stewart Morrison [a,d], Felipe Ochoa [a]

[a] School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia
[b] National ICT Australia (NICTA), Australian Technology Park, Sydney, Australia
[c] Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
[d] Department of Linguistics, University of Alberta, Edmonton, Canada

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form 26 May 2015
Accepted 7 June 2015

Keywords:
Forensic voice comparison
Likelihood ratio
Validity
Reliability
Mismatch compensation

ABSTRACT

The new paradigm for the evaluation of the strength of forensic evidence includes: The use of the likelihood-ratio framework. The use of relevant data, quantitative measurements, and statistical models. Empirical testing of validity and reliability under conditions reflecting those of the case under investigation. Transparency as to decisions made and procedures employed. The present paper illustrates the use of the new paradigm to evaluate strength of evidence under conditions reflecting those of a real forensic-voice-comparison case. The offender recording was from a landline telephone system, had background office noise, and was saved in a compressed format. The suspect recording included substantial reverberation and ventilation system noise, and was saved in a different compressed format. The present paper includes descriptions of the selection of the relevant hypotheses, sampling of data from the relevant population, simulation of suspect and offender recording conditions, and acoustic measurement and statistical modelling procedures. The present paper also explores the use of different techniques to compensate for the mismatch in recording conditions. It also examines how system performance would have differed had the suspect recording been of better quality.

1. Introduction

In *Daubert v Merrell Dow Pharmaceuticals* [1993, 509 US 579] the United States Supreme Court instructed judges to consider several factors in determining the admissibility of forensic evidence, including whether the methodology applied is scientifically valid and whether it has been empirically tested and found to have an acceptable error rate. Saks and Koehler [1] described a paradigm shift in forensic science which they proposed was in part driven by the *Daubert* ruling and in part by the shift already having occurred for DNA evidence. They "envision[ed] a paradigm shift in the traditional forensic identification sciences in which untested assumptions and semi-informed guesswork are replaced by a sound scientific foundation and justifiable protocols." (p. 895). They also proposed that "the time is ripe for the traditional forensic sciences to replace antiquated assumptions of uniqueness and perfection with a more defensible empirical and probabilistic foundation." (p. 895). The 2009 National Research Council (NRC) report to the U.S. Congress [2] was highly critical of contemporary practice across a broad range of forensic science disciplines. Their recommendations included that procedures be adopted which include "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23), "the reporting of a measurement with an interval that has a high probability of containing the true value" (p. 121), and "the conducting of validation studies of the performance of a forensic procedure" (p. 121). In response to the *R v T* ruling by the Court of Appeal of England & Wales (*R v T* [2010] EWCA Crim 2439, [2011] 1 Cr App R 9), a large number of individuals and organisations have affirmed or reaffirmed that the likelihood-ratio framework is the logically correct framework for the evaluation of forensic evidence [3–13] (see also [14–17]). The need for transparency was also a major theme in the *R v T* ruling itself and in several of the responses.

Drawing on the ongoing changes and calls for change in forensic science, Morrison and colleagues have formulated a description of a new paradigm for the evaluation of forensic evidence which includes the following key elements:

- use of the likelihood-ratio framework for the evaluation of the strength of forensic evidence
- use of approaches based on relevant data, quantitative measurements, and statistical models (relevant data is representative of the relevant population)

* Corresponding author at: School of Electrical Engineering & Telecommunications, University of New South Wales, UNSW Sydney, NSW 2052, Australia.
E-mail address: ewald.enzinger@oeaw.ac.at (E. Enzinger).

http://dx.doi.org/10.1016/j.scijus.2015.06.005
1355-0306/

- empirical testing of the validity and reliability of the forensic analysis system under conditions reflecting those of the case under investigation.

For the first time here we propose the promotion of a fourth concern to be an explicit member of this list of key elements:

- transparent reporting of choices made and procedures employed.

An early formulation of Morrison and colleagues' conception of the new paradigm, and a description of the history of the paradigm shift in forensic voice comparison appeared in Morrison [18]. Another early formulation appeared in Morrison [19], and later formulations in Morrison [9], Morrison [20], and Morrison and Stoel [21]. Morrison et al. [22] focussed particularly on the selection of the relevant population for the defence hypothesis, and Morrison [23] on procedures for empirically testing validity and reliability within the likelihood-ratio framework.

1.1. Description of general procedures for implementation of the new paradigm

The following is a description of general procedures for performing a source-level forensic comparison within the new paradigm (it is based on the description in Morrison and Stoel [21]):

First, the forensic scientist must define and communicate the prosecution and defence hypotheses as they understand them. A forensic likelihood ratio is the answer to a specific question,[1] and to make sense of the likelihood ratio both the forensic scientist and the trier of fact need to understand that question. The question is specified by two hypotheses: the prosecution hypothesis, which pertains to the numerator of the likelihood ratio, and the defence hypothesis, which pertains to the denominator. A typical prosecution hypothesis is that the sample of questioned origin comes from the same source as the sample of known origin. A typical defence hypothesis is that the sample of questioned origin does not come from the same source as the sample of known origin, but from some other source in the relevant population. The relevant population is specific to the particular case under investigation (see, for example, Curran et al. [24], on glass and Kerkhoff et al. [25], on firearms). In most jurisdictions, it is not common for the court to provide the forensic scientist with explicit hypotheses to test prior to the forensic scientist beginning their analysis. In such circumstances, the forensic scientist must therefore use their own judgement and adopt hypotheses which they believe will be of interest to the trier of fact. Analysis cannot proceed unless both a prosecution and defence hypothesis are either provided to or adopted by the forensic scientist. A legitimate question to debate before the trier of fact would be whether the alternative hypothesis adopted by the forensic scientist is appropriate. That is, does it lead to a likelihood ratio which answers the question that the trier of fact wants to have answered.[2] By making their adopted hypotheses explicit, the forensic scientist facilitates consideration of this important question.

Next, the forensic scientist must obtain a sample from the relevant population. This sample is to be used to train the model which will calculate the denominator of the likelihood ratio. A legitimate issue to debate before the trier of fact would be whether the sample is sufficiently representative of the relevant population (see Hancock et al. [27], Morrison [9]).

The forensic scientist must make measurements which quantity the properties of the sample of known origin (suspect sample), the sample of questioned origin (offender sample), and each item in the sample representative of the relevant population. These measurements constitute relevant data.

Next, the forensic scientist must choose the statistical models that they will use to calculate the likelihood ratio. Part of the expertise of the forensic scientist is to select a model which they expect will give a reasonable approximation of the distribution of the population without overfitting the model to the particular training data. They can conduct tests using development data to help them select a model which gives what they themselves consider to be sufficiently acceptable performance under the conditions of the case under investigation.[3] The models should be trained and optimised using data which reflect the conditions of the case under investigation. In a forensic-voice-comparison case this would include recording and transmission channel (e.g., landline or mobile telephone, compression algorithms), background noise, reverberation, speaking style (conversation, formal speech), etc. To avoid condition-dependent bias in the calculation of the denominator versus the numerator of the likelihood ratio, the data used to train the model for the denominator should be in the same condition as the known-origin data which is used to train the model for the numerator. Ideally the statistical models would also incorporate techniques which attempt to compensate for mismatches between the conditions of samples of known and questioned origin. The description of the conditions also forms part of the specific question which is to be answered by the likelihood ratio. For example: What is the probability of getting the properties of the distorted partial latent mark if it were produced by the same finger as made the high-quality suspect fingerprint versus if it were made by a finger of someone else from the relevant population? The forensic scientist should communicate to the trier of fact the conditions of the case as they understand them, and how they form part of the specific question to be answered by the likelihood ratio.

Once relevant training data have been selected and a model has been chosen, trained, and optimised to the conditions of the case under investigation, the system should be frozen, i.e., no other changes are allowed. Then the system should be tested using new pairs of sample items drawn from the relevant population and reflecting the conditions of the actual samples of known and questioned origin from the case under investigation. In this way the forensic scientist obtains an indication of how well the system is expected to perform on previously unseen data from the relevant population under these conditions. Testing using samples from some other population or under different conditions will not be informative as to how well the system is expected to perform on the actual samples of known and questioned origin from the case under investigation. Testing using some other population and/or under some other conditions, could potentially be highly misleading with respect to the performance of the system in the particular case under investigation. An issue for debate would be whether the conditions of the training and test data adequately reflect the conditions of the samples of known and questioned origin.

If the judge at an admissibility hearing or the trier of fact at trial is satisfied that the samples adequately reflect the relevant population and conditions specific to the case, and is satisfied that the model is answering a question which is relevant to the trier of fact, then they should consider whether the empirically demonstrated degree of validity and reliability of the system is sufficient for the output to be of use to the trier of fact. If they are not satisfied on any of these points, then the output of the system will be of little or no value to them. It is therefore essential that the forensic scientist be transparent as to what they have done, and that they present the results of validity and reliability testing.

After the performance of the system has been empirically tested, the system and the test results are frozen, i.e., no other changes are allowed to the system, and the test data cannot be changed and new tests cannot be run. The last thing the forensic scientist does as part of the analysis is

[1] The specific question in the present case is provided in Section 2.1.

[2] As pointed out in Rose [26, p. 65], in addition to the likelihood ratio that the forensic scientist calculates being dependent on the choice of population, the prior odds adopted by the trier of fact may also be dependent on the choice of population.

[3] Ultimately, it will be up to the trier of fact to decide whether what the forensic scientist has done is acceptable, but in the first instance the forensic scientist themself must be satisfied with what they have done.