A New Paradigm for Forensic Science

Geoffrey Stewart Morrison

Ewald Enzinger

\[
\frac{p(E|H_p)}{p(E|H_d)}
\]
Abstract

- In Europe there has been a great deal of concern about the logically correct way to evaluate the strength of forensic evidence. The 2015 European Network of Forensic Science Institutes’ Guideline for Evaluative Reporting in Forensic Science recommends the use of the likelihood-ratio framework. In the United States there has been a great deal of concern about the validity and reliability of forensic science, as stressed in the 2009 National Research Council Report to Congress. In England & Wales the Forensic Science Regulator’s 2014 Codes of Practice and Conduct require validation of methods in all branches of forensic science to be demonstrated within the next few years. Additional current concerns include the need for transparency, as expressed in the 2010 England & Wales Court of Appeal ruling in *R v T* and the multiple published responses to that ruling, and the need to adopt procedures which minimise the potential for cognitive bias, as expressed in the 2012 US National Institute of Science and Technology and National Institute of Justice report on human factors in latent fingerprint analysis.

- For a number of years, the presenter and his colleagues have been developing a description of a paradigm for the evaluation of forensic evidence which addresses the concerns from both sides of the Atlantic, and which is applicable across all branches of forensic science. The paradigm includes: the calculation of likelihood ratios using relevant data, quantitative measurements, and statistical models; and empirical testing of the validity and reliability of forensic analysis systems under conditions reflecting those of the case under investigation. This presentation provides an introduction to the paradigm and explains how it addresses current concerns. If time permits, the presentation will also include a description of the implementation of the paradigm in a real forensic case. The example case is a forensic voice comparison case, but the principles exemplified are applicable across all branches of forensic science. The presenter has an interactive presentation style, adapts the presentation to the audience on the day, and encourages audience members to ask questions as the presentation goes along.
Some people attending the OSAC meeting

William C Thompson

James M Curran

Reva Schwarz

Bryan Found
CONCERNS

• **Logically correct framework for evaluation of forensic evidence**
  – ENFSI Guideline for Evaluative Reporting 2015

• **But what is the warrant for the opinion expressed? Where do the numbers come from?**
  – Risinger at ICFIS 2011

• **Demonstrate validity and reliability**

• **Transparency**
  – *R v T* 2010 and responses

• **Reduce potential for cognitive bias**
  – NIST/NIJ Human Factors in Latent Fingerprint Analysis 2012
PARADIGM

• **Use of likelihood ratio framework**
  
  • Logically correct framework for evaluation of evidence.
  
  • Specific prosecution and defence hypotheses adopted by forensic scientist must be explained to judge at admissibility hearing / trier of fact at trial.

  probability of obtaining the acoustic properties on the offender recording if it were produced by the suspect versus if it were produced by some other speaker selected at random from the relevant population

  • Is the question appropriate?

  • Question must be understood in order to understand answer.
Likelihood ratios

- Adopted as standard for evaluation of DNA evidence in mid 1990’s
Likelihood ratios

- Association of Forensic Science Providers (2009)
  - Standards for the formulation of evaluative forensic science expert opinion

- 31 signatories [from Aitken to Zadora] (2011)
  - Expressing evaluative opinions: A position statement

  - Guideline for evaluative reporting in forensic science
Likelihood ratios

- **Robertson & Vignaux (1995)**

- **Balding & Steele (2015)**

- **Morrison (2010)**

- **Morrison (2012)**

- **Jackson (2009)**
PARADIGM

• **Calculation of numeric likelihood ratios using relevant data, quantitative measurements, and statistical models**

  • Sample from the relevant population specified in the defence hypothesis. Sufficiently representative?

  • Data reflective of conditions of suspect and offender samples. Sufficiently reflective?

  • Report output of statistical model, keep subjective elements far from the conclusion. Do not report conclusions which are primarily or directly based on subjective judgement.
Calculate and report numeric likelihood ratios

- Morrison & Stoel (2014)

Cognitive bias

- Found (2015)
PARADIGM

- **Empirical testing of validity and reliability under conditions reflecting those of the case under investigation**

  - Performance under different conditions may be very different.
  
  - Sample from the relevant population specified in the defence hypothesis. Sufficiently representative?
  
  - Data reflective of conditions of suspect and offender samples. Sufficiently reflective?
  
  - Are the number of test trials sufficient?
  
  - Test the system actually employed, including human operator.
  
  - Metrics of system performance should be compatible with the likelihood ratio framework.
Testing should be method agnostic
Testing should be method agnostic

1024

42

1,000,000

To be, or not to be
Validity and Reliability

- The National Research Council report to Congress on *Strengthening Forensic Science in the United States* (2009) urged 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)

  - “the conducting of validation studies of the performance of a forensic procedure” (p. 121)
Validity and Reliability

The Forensic Science Regulator of England & Wales’ Codes of Practice and Conduct (2014) require:

- “all technical methods and procedures used by a provider shall be validated.” (§20.1.1)

- “Even where a method is considered standard and is in widespread use, validation will still need to be demonstrated.” (§20.1.3)

- “validation shall be carried out using simulated casework material ... and ... where appropriate, with actual casework material” (§20.7.3)

- “demonstrate that they can provide consistent, reproducible, valid and reliable results” (§20.9.1)
Validity and Reliability

- **Morrison (2014)**

- **Morrison (2011)**
Examples of casework conducted within the new paradigm

- **Enzinger et al (2016)**

- **Enzinger & Morrison (2015)**
Real Case

• **Offender recording**
  Telephone call made to a financial institution’s call centre
  – landline
  – call centre background noise (babble, typing)
  – saved in a compressed format
  – 46 seconds net speech

• **Suspect recording**
  Police interview
  – reverberation
  – ventilation system noise
  – saved in a compressed format
Strict chronological order for analysis

- Determine prosecution and defence hypotheses to adopt
  - includes defining the relevant population

- Obtain data representative of the relevant population, and reflective of the conditions of the suspect and offender recordings
  - split these into training data and test data

- Train a forensic voice comparison system

- Test the performance of the forensic voice comparison system

- Calculate a likelihood ratio for the comparison of the suspect and offender recording

Document all decisions made, actions taken, and results obtained at each stage.

Do not at any time move back to an earlier stage.
Defence hypothesis and relevant population adopted

- Relevant population chosen based on offender recording
  Obvious that the speaker was
  - adult male
  - speaking Australian English

- We had previously invested in collecting a database of voice recordings
  which included:
  - 231 adult male Australian English speakers
  - high-quality recordings
  - speaking styles:
    - information exchange over the telephone
    - simulated police interview
  - multiple non-contemporaneous recordings in each speaking style
Simulation of offender-recording conditions

\[ x_r[i] \rightarrow \downarrow 8kHz \rightarrow [300 \text{ Hz} \quad 3400 \text{ Hz}] \rightarrow \text{a-Law compression/decompression} \]

\[ \sqrt{e_s \over e_r} \rightarrow y_r[i] \]

\[ x_n[i] \rightarrow \text{offender recording noise} \]

play audio
Simulation of suspect-recording conditions

\[ x_r[i] \rightarrow \text{MPEG-1 layer 2 compression/decompression} \]

\[ \text{scaling } \alpha = \sqrt{\frac{e_s}{e_r}} \rightarrow y_r[i] \]

\[ x_n[i] \rightarrow \text{suspect recording noise} \]
Selection of samples representative of the relevant population

- We were only asked to compare the suspect and offender recordings because a police officer had listened to them and thought they were sufficiently similar sounding that it was worth submitting them for forensic analysis.

- Listeners similar to the police officer selected the speakers from the database to include in the sample of the relevant population:
  - same gender
  - approximately the same age
  - same linguistic background (monolingual Australian English speakers)

- Listened to offender recording and to suspect-condition database recordings.
Selection of samples representative of the relevant population

Does the speaker on recording B sound sufficiently similar to the speaker on recording A (offender recording) that you would submit recording B for forensic comparison with recording A?
### Selection of samples representative of the relevant population

The number of speakers selected by $N$ or more listeners.

<table>
<thead>
<tr>
<th>number of listeners, $N$</th>
<th>11</th>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of speakers selected by $N$ or more listeners</td>
<td>16</td>
<td>24</td>
<td>34</td>
<td>42</td>
<td>51</td>
<td>75</td>
<td>100</td>
<td>128</td>
<td><strong>166</strong></td>
<td>195</td>
<td>216</td>
</tr>
</tbody>
</table>

- Training data: 423 recordings from 105 speakers
- Test data: 222 recordings from 61 speakers
- Test protocol included 9669 comparison pairs
Quantitative acoustic measurements

- mel frequency cepstral coefficients + deltas

- Suspect-condition and offender-conditions recordings made same durations as the actual suspect and offender recordings (in MFCC frames)
Quantitative acoustic measurements
Statistical models

- **GMM-UBM**
  - suspect model trained using **suspect** data
  - population model (UBM) trained using **suspect-condition** data from **sample of the population**
  - same mismatch with **offender** data

- Score to likelihood ratio conversion (logistic regression)
  - trained using pairs of recordings from **sample of the population**, one in **suspect condition**, the other in **offender condition**

- System also included mismatch compensation techniques (feature warping, probabilistic feature mapping)
Test results

- $C_{llr}$-pooled: 0.423
- $C_{llr}$-mean: 0.344
- 95% CI: ±0.95

Test data:
- pairs of recordings from sample of the population, one in suspect condition, the other in offender condition
Comparison of suspect and offender recordings

- LR: 343
- \( \log_{10} \) LR: 2.54
- 98% CI: ±1.13 [25 .. 4599]
- Probability of equal or stronger misleading evidence: 0.00033
Conclusions

- Based on our calculations we estimate that the probability of obtaining the acoustic properties of the speech on the offender recording is approximately 300 times greater had it been produced by the defendant than had it been produced by some other speaker selected at random from the relevant population.

- Our best estimate for the strength of the evidence is a likelihood ratio of 343, and based on tests of our system we are 99% certain that the probability of obtaining the acoustic properties of the offender sample is at least 25 times greater had it been produced by the defendant than had it been produced by some other speaker selected at random from the relevant population.

- Based on tests of our system, we estimated that the probability of observing a likelihood ratio of equal to or greater than 343 if the offender sample were produced by a speaker selected at random from the relevant population is 0.00033, less than four in ten thousand.
INTERPOL Survey

• Number of responses:
  - 91 from 69 countries
  - 44 had speaker identification capabilities in house or via external laboratories
INTERPOL Survey

- Identification / exclusion / inconclusive
- Numeric posterior probability
- Numeric likelihood ratio
- Verbal posterior probability
- Verbal likelihood ratio
- UK framework

Count - Global

0 5 10 15 20 25
Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case

http://databases.forensic-voice-comparison.net/#forensic_eval_01
Thank You

http://geoff-morrison.net/

http://forensic-evaluation.net/