A New Paradigm for the Evaluation of Forensic Evidence

and its implementation in forensic voice comparison

Geoffrey Stewart Morrison

Ewald Enzinger

$$\frac{p(E|H_p)}{p(E|H_d)}$$

Abstract & Biographies

- 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 minimize 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 presenters and their 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 overview of the paradigm and 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 also applicable in other branches of forensic science.
- Dr Morrison is an independent forensic consultant based in Vancouver, British Columbia, Canada. He is an Adjunct Associate Professor at the Department of Linguistics, University of Alberta; and currently a Guest Editor for the journal *Science & Justice*. He has previously been Scientific Counsel, Office of Legal Affairs, INTERPOL General Secretariat; Director of the Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales; a Subject Editor for the journal *Speech Communication*; and Chair of the Forensic Acoustics Subcommittee of the Acoustical Society of America. He has authored more than 50 refereed and invited academic publications, more than 30 in forensic science, and has conducted research in collaboration with police services in Australia and in Europe. He has worked on forensic casework in Australia and in the United States, and has worked at the behest of both the prosecution and the defense. In 2015 he advised defense counsel in a US Federal Court *Daubert* hearing on the admissibility of a forensic voice comparison analysis proffered by the prosecution.
- Dr Enzinger is an independent forensic consultant based in Corvallis, Oregon, United States of America. He is also currently a Guest Editor for the journal *Speech Communication*. He previously held research appointments at the Acoustics Research Institute of the Austrian Academy of Sciences, and recently completed a PhD at the School of Electrical Engineering & Telecommunications, University of New South Wales. He has worked on research projects in collaboration with federal and state police services in Australia and in collaboration with the Germany Federal Police. His doctoral dissertation *Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence* has been described as "a remarkable and significant step forward in the field ... [that] represents one of the strongest research advancements in this domain to date."

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**
  – *Daubert* 1993; National Research Council Report 2009; Forensic Science Regulator Codes of Practice 2014
- **Transparency**
  – *R v T* 2010 and responses
- **Reduce potential for cognitive bias**
  – NIST/NIJ Human Factors in Latent Fingerprint Analysis 2012
- **Communicate strength of evidence to trier of fact**

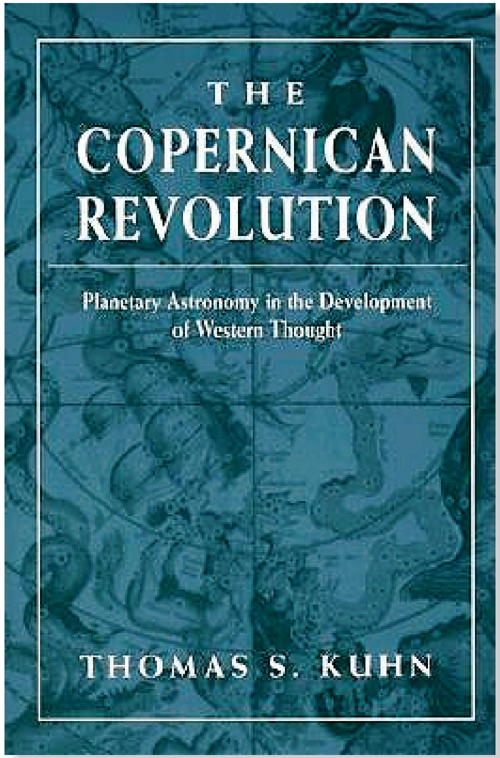
PARADIGM

- **Use of the likelihood-ratio framework for the evaluation of forensic evidence**
  - logically correct

- **Use of relevant data, quantitative measurements, and statistical models**
  - transparent and replicable
  - robust to cognitive bias

- **Empirical testing of validity and reliability under conditions reflecting those of the case under investigation**
  - only way to know how well it works

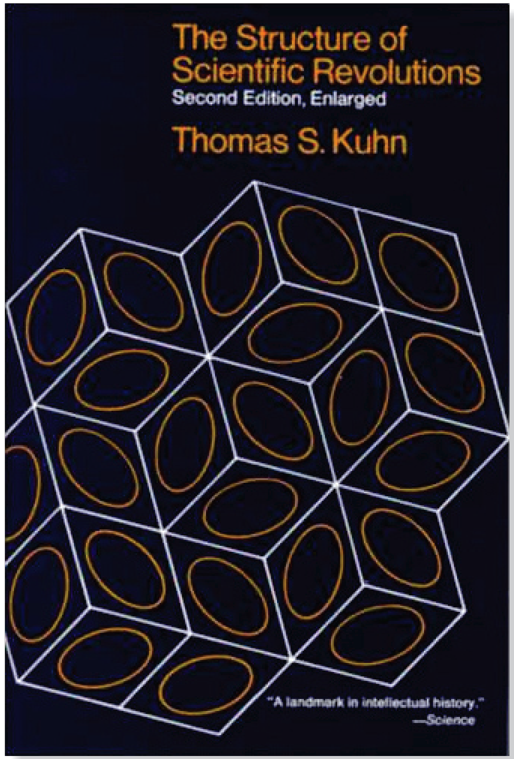
PARADIGM SHIFT



THE
COPERNICAN
REVOLUTION

Planetary Astronomy in the Development
of Western Thought

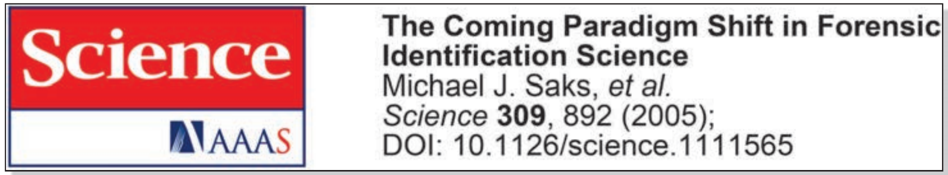
THOMAS S. KUHN



The Structure of
Scientific Revolutions
Second Edition, Enlarged

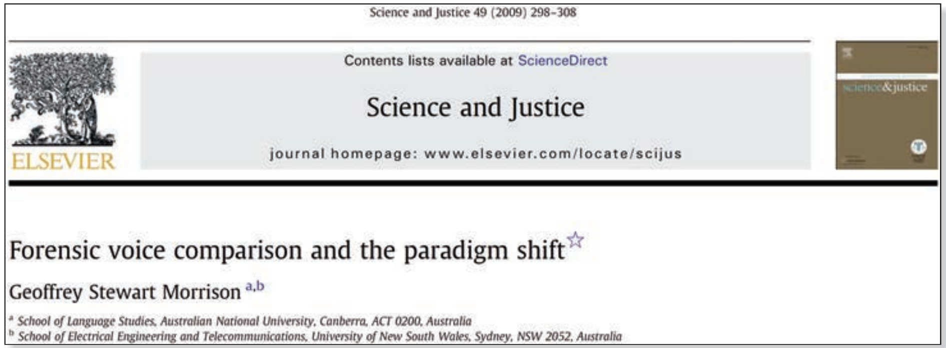
Thomas S. Kuhn

"A landmark in intellectual history."
—Science



Science
AAAS

The Coming Paradigm Shift in Forensic Identification Science
Michael J. Saks, *et al.*
Science **309**, 892 (2005);
DOI: 10.1126/science.1111565



Science and Justice 49 (2009) 298–308

Contents lists available at ScienceDirect

Science and Justice

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

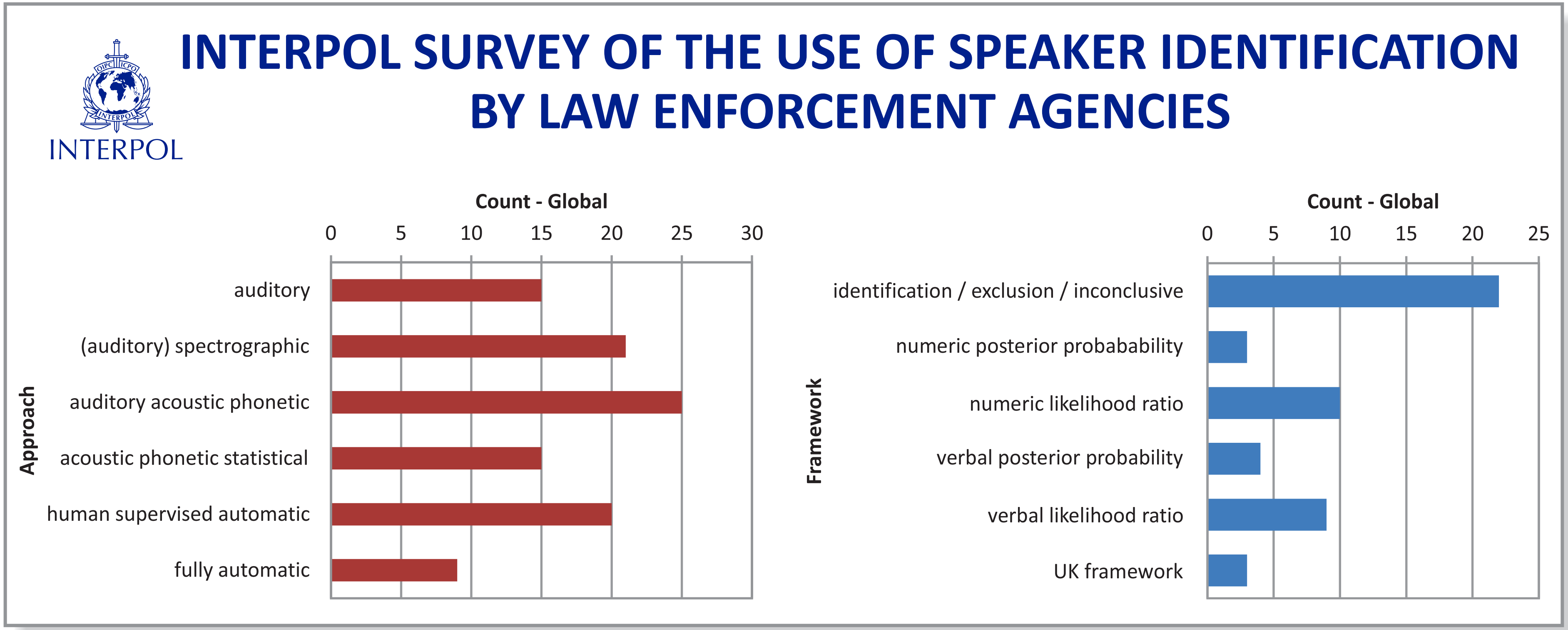
ELSEVIER

Forensic voice comparison and the paradigm shift ☆

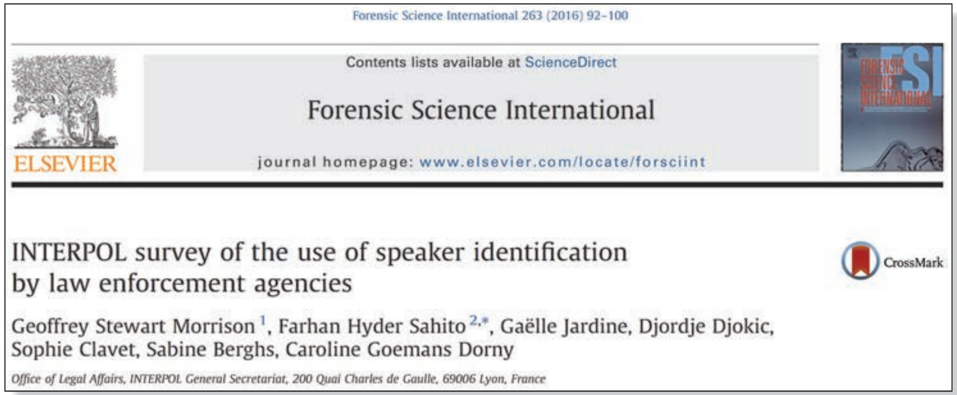
Geoffrey Stewart Morrison [a,b]

[a] School of Language Studies, Australian National University, Canberra, ACT 0200, Australia
[b] School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

PARADIGM SHIFT in Forensic Speech Science



INTERPOL SURVEY OF THE USE OF SPEAKER IDENTIFICATION BY LAW ENFORCEMENT AGENCIES



Forensic Science International 263 (2016) 92–100

Contents lists available at ScienceDirect

Forensic Science International

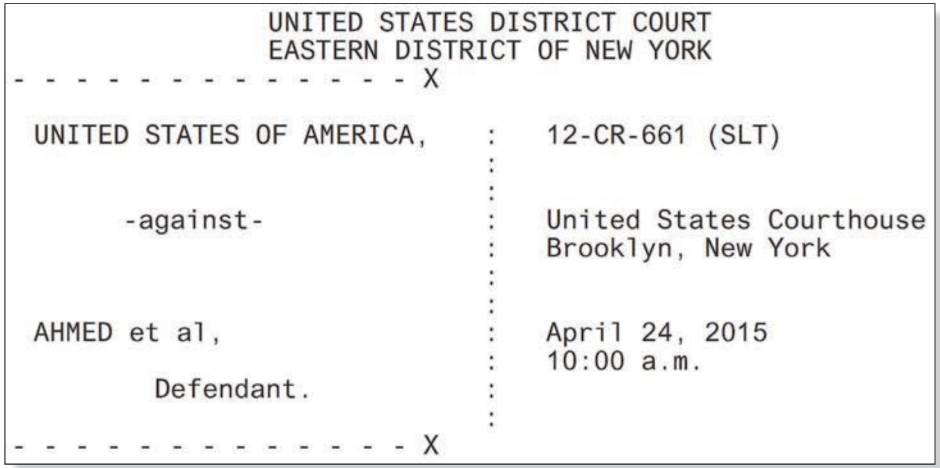
journal homepage: www.elsevier.com/locate/forsciint

ELSEVIER
CrossMark

INTERPOL survey of the use of speaker identification by law enforcement agencies

Geoffrey Stewart Morrison[1], Farhan Hyder Sahito[2,*], Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs, Caroline Goemans Dorny

Office of Legal Affairs, INTERPOL General Secretariat, 200 Quai Charles de Gaulle, 69006 Lyon, France



UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF NEW YORK

- - - - - - - - - - - - - - - X

UNITED STATES OF AMERICA, : 12-CR-661 (SLT)

-against- : United States Courthouse
Brooklyn, New York

AHMED et al, : April 24, 2015
10:00 a.m.

Defendant.

- - - - - - - - - - - - - - - X

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

- The hair at the crime scene is blond
- The suspect has blond hair

Likelihood ratios

- The hair at the crime scene is blond
- The suspect has blond hair

- Does that mean that the suspect and offender are the same person?
- Does it mean that it is highly probable that the suspect and offender are the same person?

Likelihood ratios

- The hair at the crime scene is blond
- The suspect has blond hair

$$\frac{p(\text{ blond hair at crime scene } \mid \text{ suspect is source})}{p(\text{ blond hair at crime scene } \mid \text{ someone else is source})}$$

- Someone else selected at random from the relevant population
- What is the relevant population?

Likelihood ratios

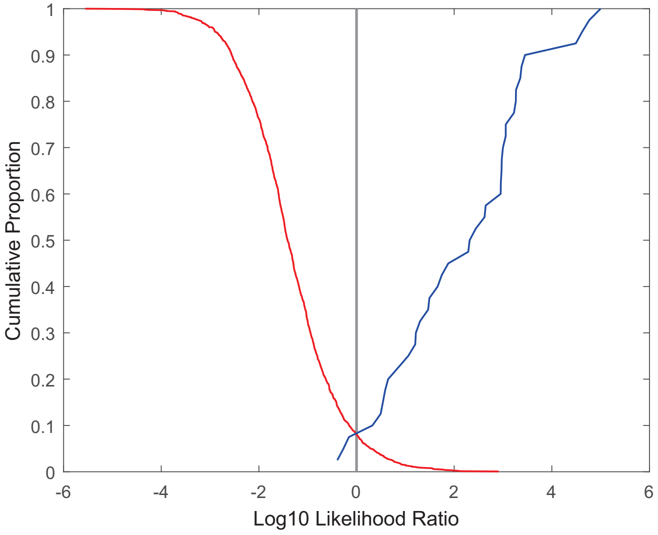
- The hair at the crime scene is blond
- The suspect has blond hair

$$\frac{p(\text{ blond hair at crime scene} \mid \text{suspect is source})}{p(\text{ blond hair at crime scene} \mid \text{someone else is source})}$$

- Someone else selected at random from the relevant population
- What is the relevant population?
    - Stockholm
    - Beijing
    - Orlando

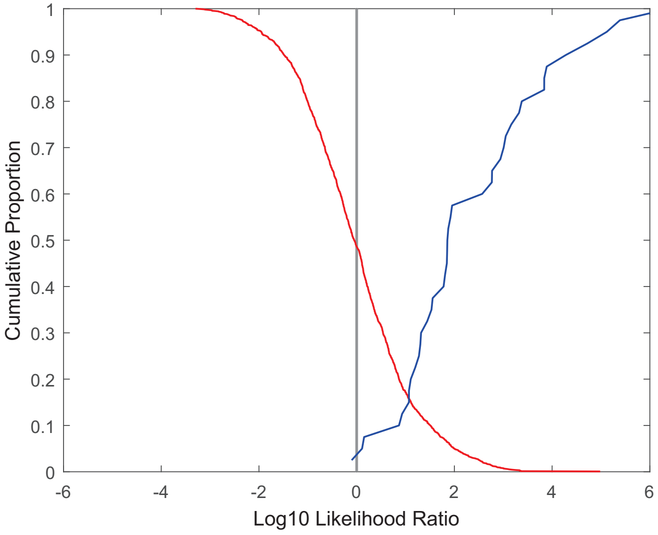
Likelihood ratios

wrong training set
wrong test set

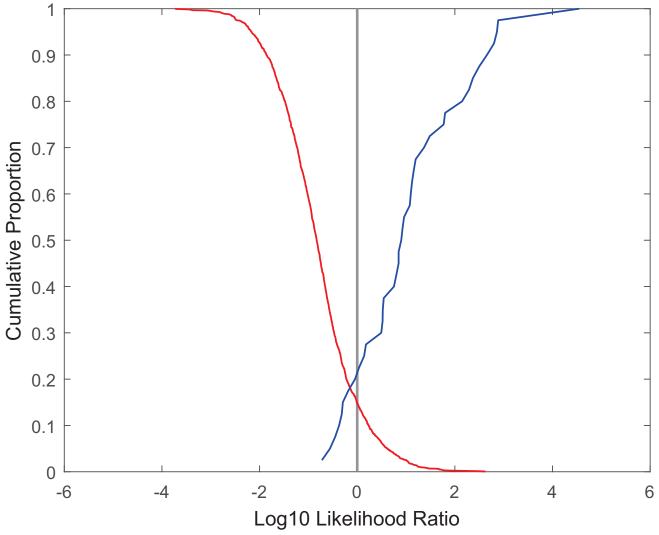


95% Australian English
5% Mandarin Chinese

wrong training set
right test set



right training set
right test set



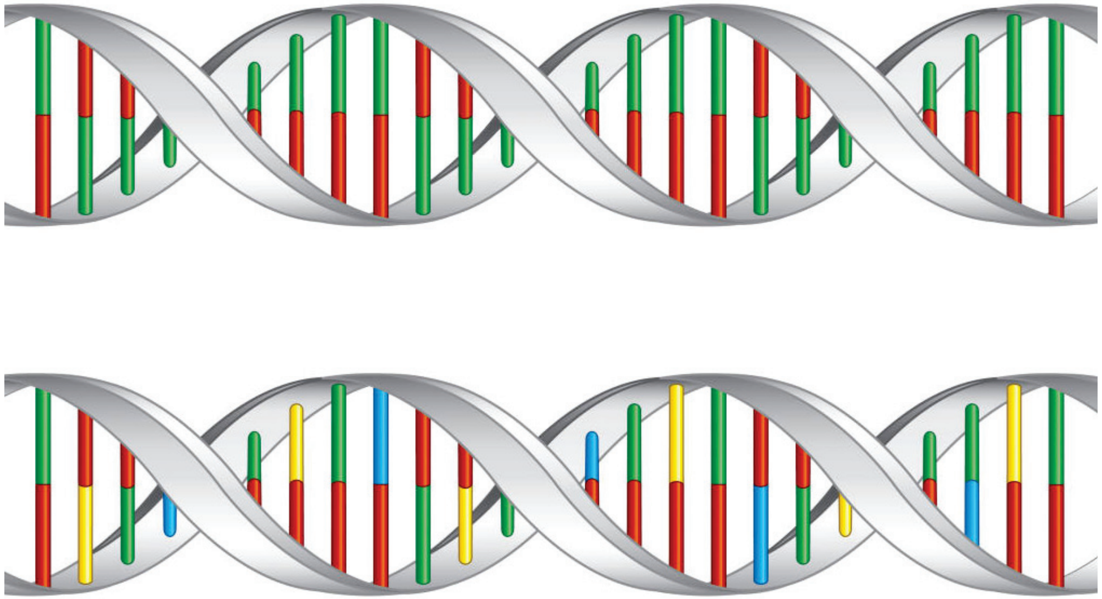
100% Mandarin Chinese

Refining the relevant population in forensic voice comparison – A response to Hicks *et alii* (2015
The importance of distinguishing information from evidence/observations when formulating propositions

Geoffrey Stewart Morrison[a,b,*], Ewald Enzinger[a], Cuiling Zhang[c,d]

Likelihood ratios

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



Likelihood ratios

AFSP
ASSOCIATION OF FORENSIC SCIENCE PROVIDERS

- 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*

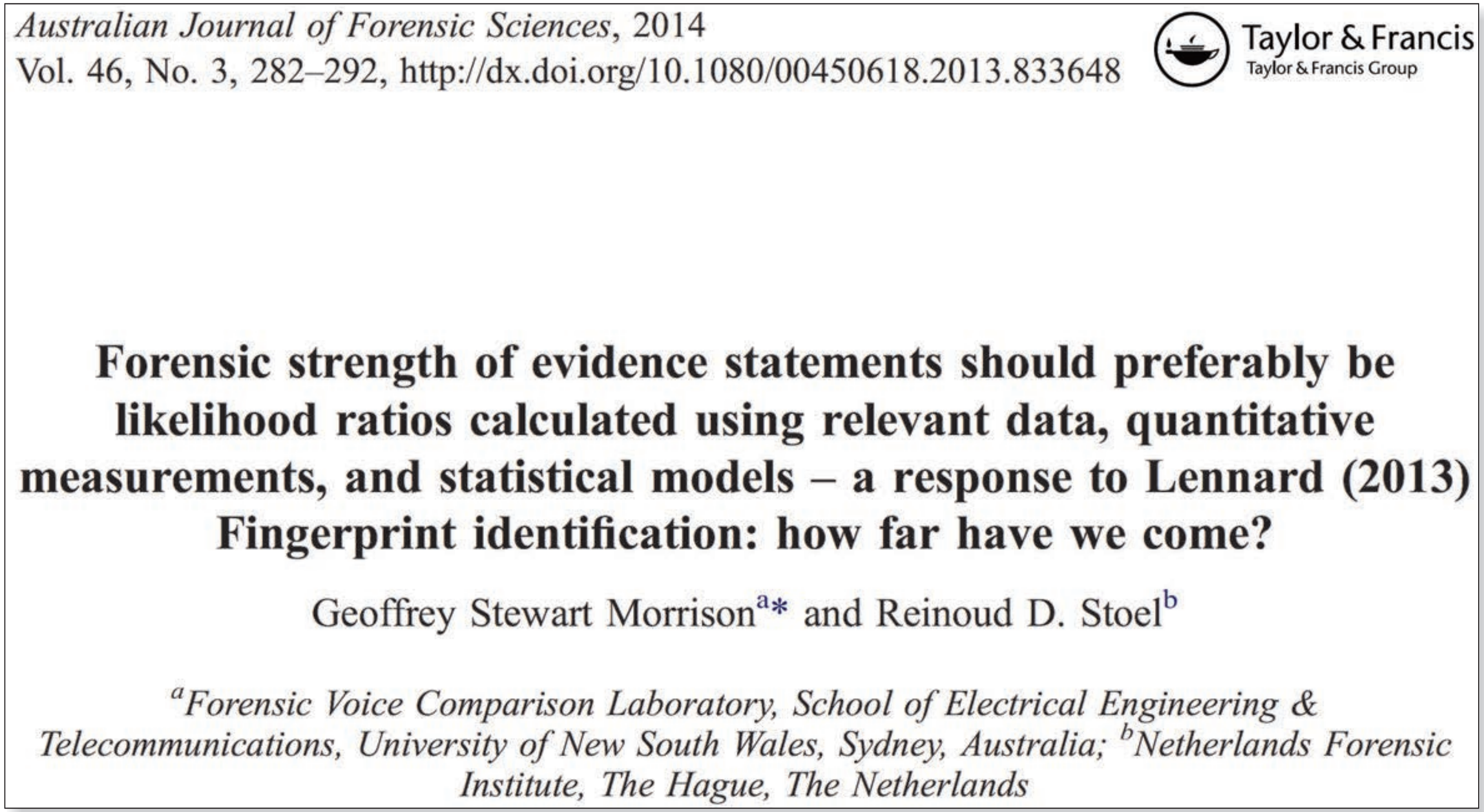
- European Network of Forensic Science Institutes (2015)
  - *Guideline for evaluative reporting in forensic science*

ENFSI

PARADIGM

- **Calculation of numeric likelihood ratios using relevant data, quantitative measurements, and statistical models**
  - Transparent and replicable
  - Robust to cognitive bias
    - 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.
  - Sample from the relevant population specified in the defence hypothesis. Sufficiently representative?
  - Data reflective of conditions of suspect and offender samples. Sufficiently reflective?

Calculation of numeric likelihood ratios



Australian Journal of Forensic Sciences, 2014
Vol. 46, No. 3, 282–292, http://dx.doi.org/10.1080/00450618.2013.833648

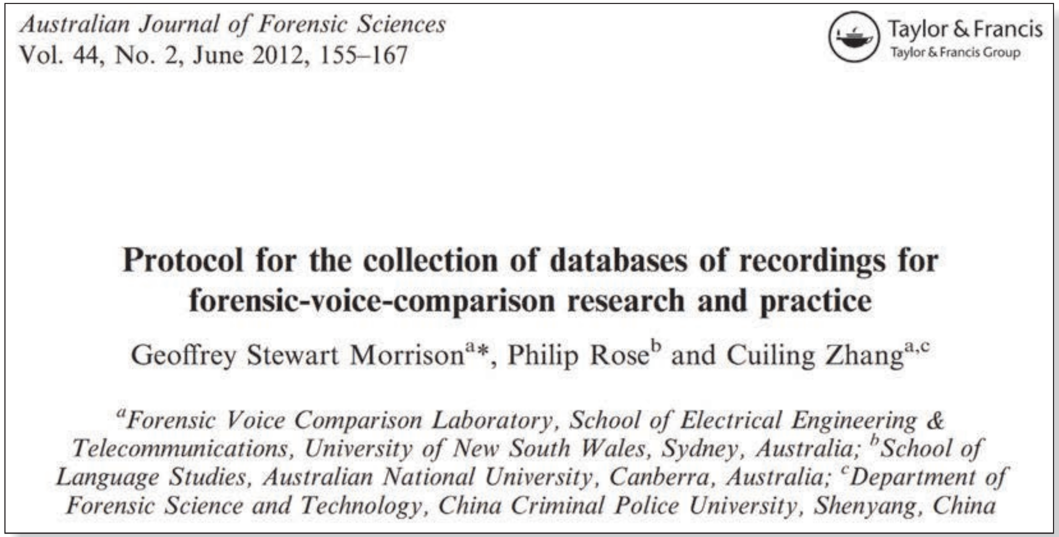
Taylor & Francis
Taylor & Francis Group

Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: how far have we come?

Geoffrey Stewart Morrison[a]* and Reinoud D. Stoel[b]

[a]*Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia;* [b]*Netherlands Forensic Institute, The Hague, The Netherlands*

Calculation of numeric likelihood ratios



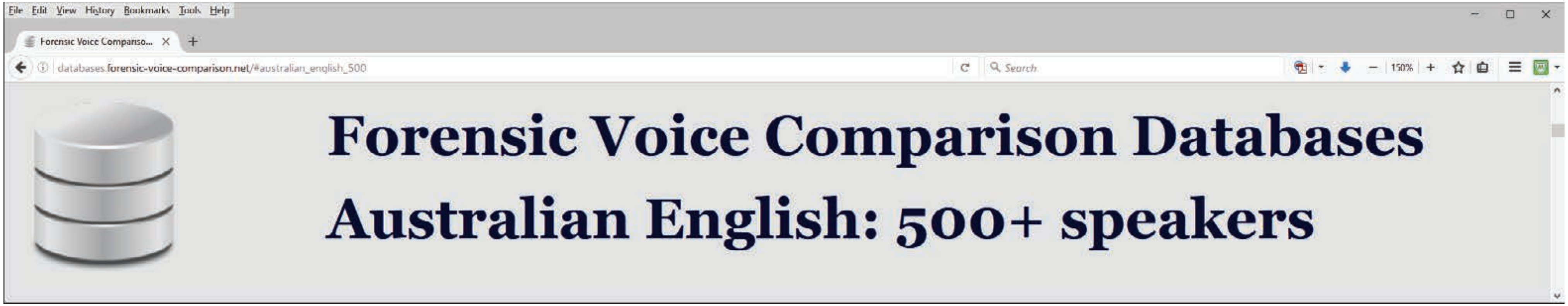
Australian Journal of Forensic Sciences
Vol. 44, No. 2, June 2012, 155–167

Taylor & Francis
Taylor & Francis Group

Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice

Geoffrey Stewart Morrison[a]*, Philip Rose[b] and Cuiling Zhang[a,c]

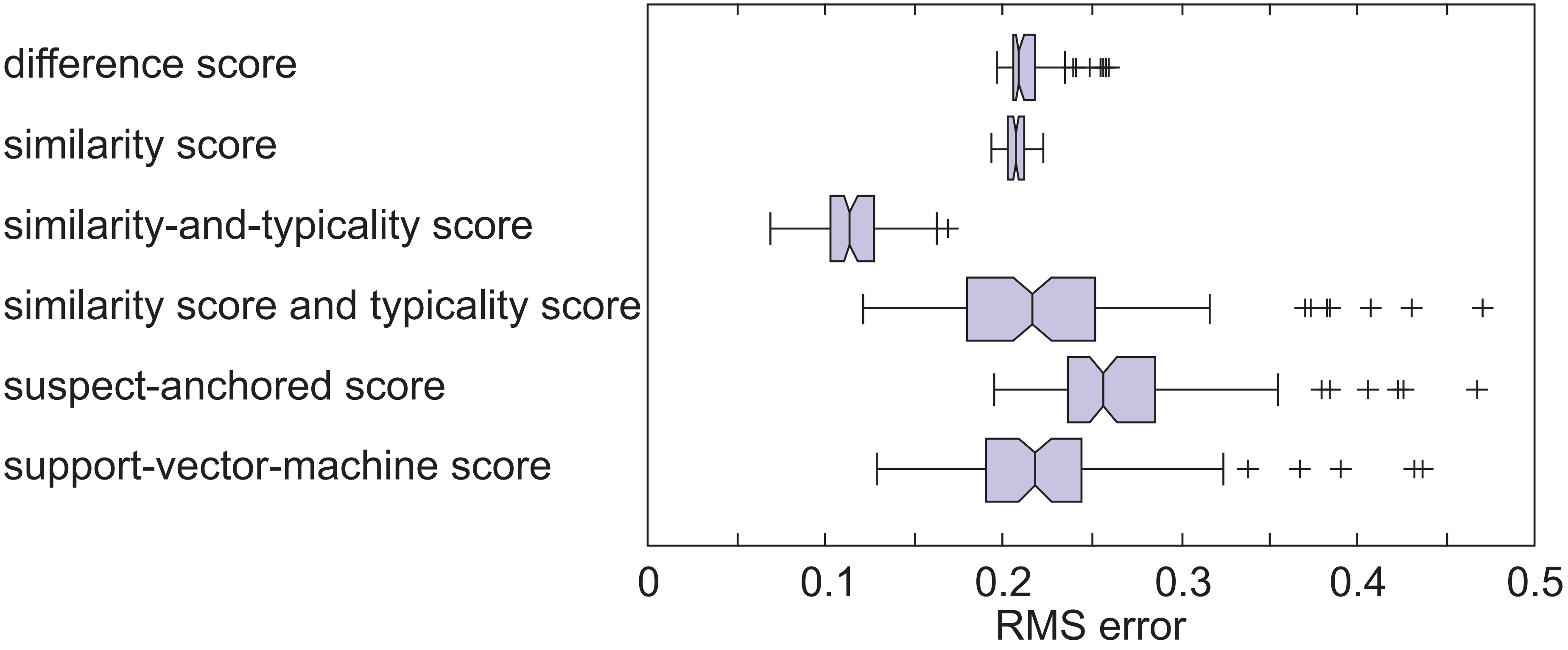
[a]*Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia;* [b]*School of Language Studies, Australian National University, Canberra, Australia;* [c]*Department of Forensic Science and Technology, China Criminal Police University, Shenyang, China*


File
Edit
View
History
Bookmarks
Tools
Help
Forensic Voice Compariso...
databases.forensic-voice-comparison.net/#australian_english_500
Search
150%

Forensic Voice Comparison Databases

Australian English: 500+ speakers

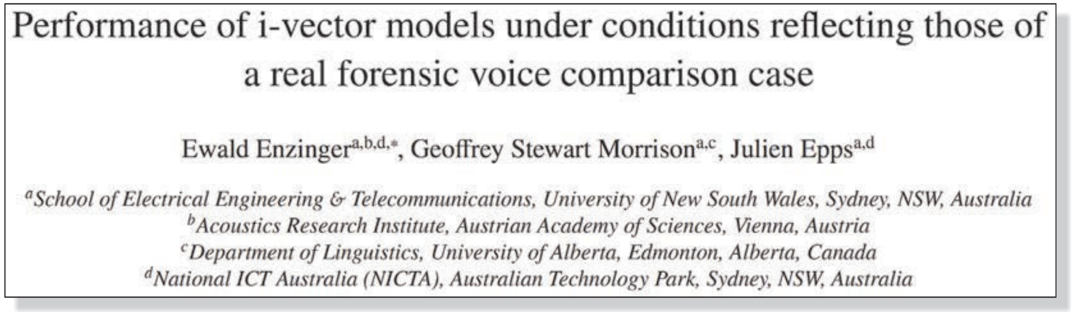
Calculation of numeric likelihood ratios



Forensic likelihood ratios
should not be based on
similarity scores or difference scores

Geoffrey Stewart Morrison
Ewald Enzinger

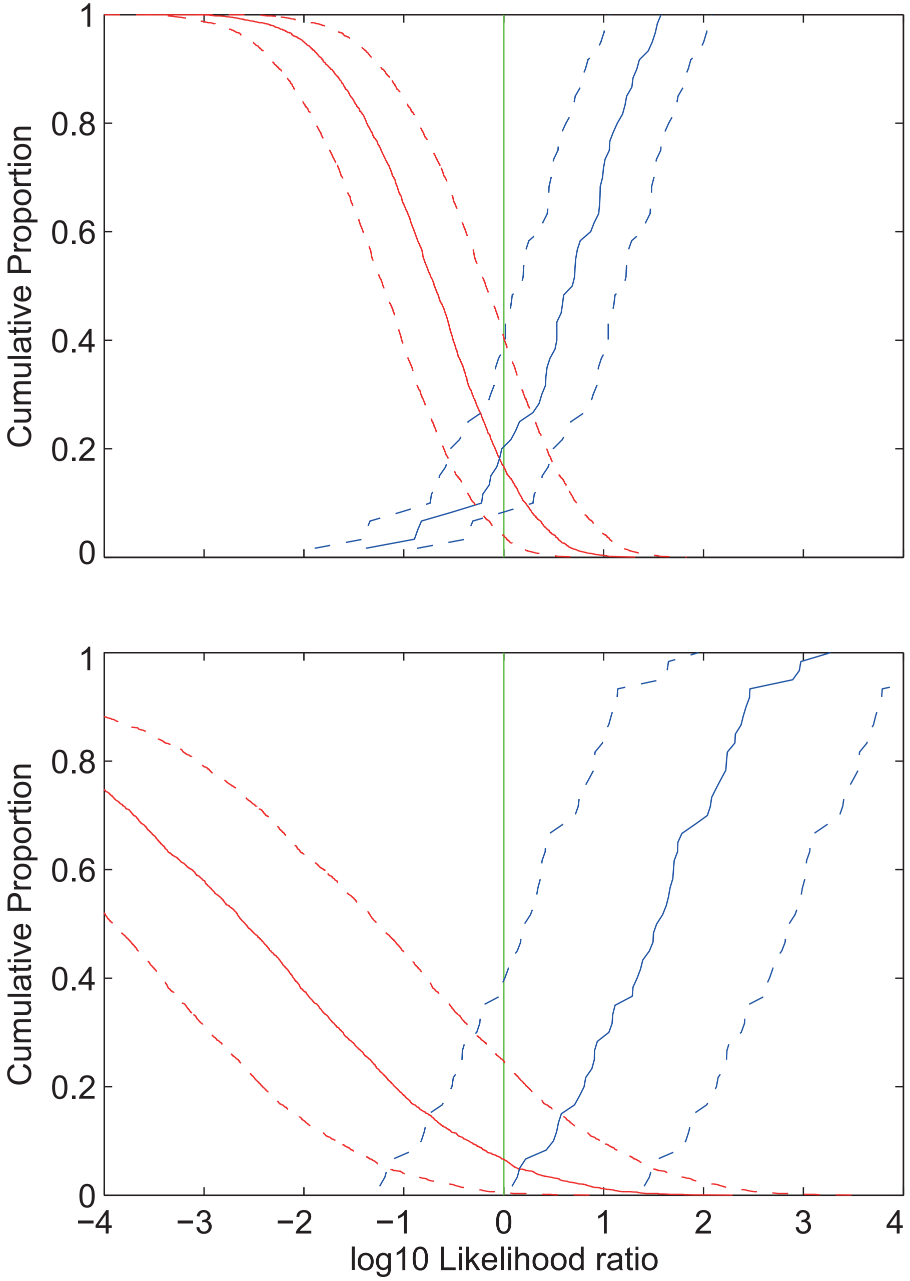
Calculation of numeric likelihood ratios



Performance of i-vector models under conditions reflecting those of a real forensic voice comparison case

Ewald Enzinger[a,b,d,*], Geoffrey Stewart Morrison[a,c], Julien Epps[a,d]

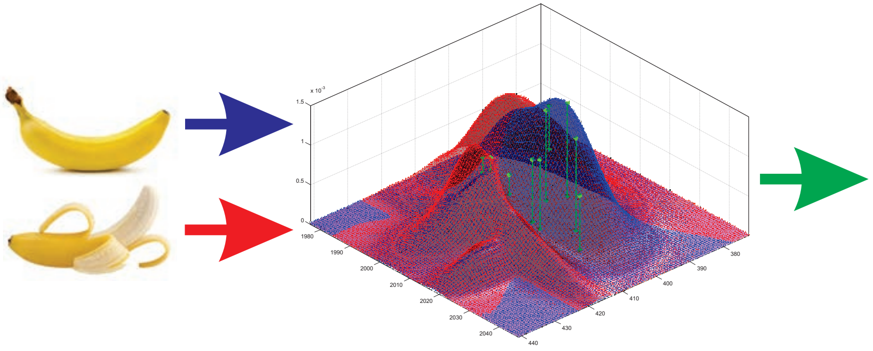
[a] *School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, NSW, Australia*
[b] *Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*
[c] *Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada*
[d] *National ICT Australia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



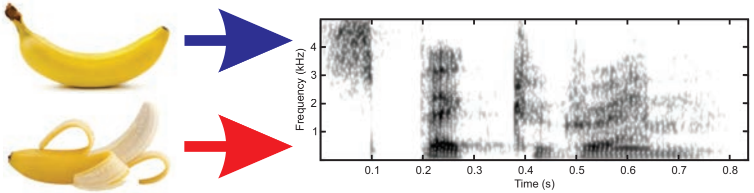
PARADIGM

- **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.

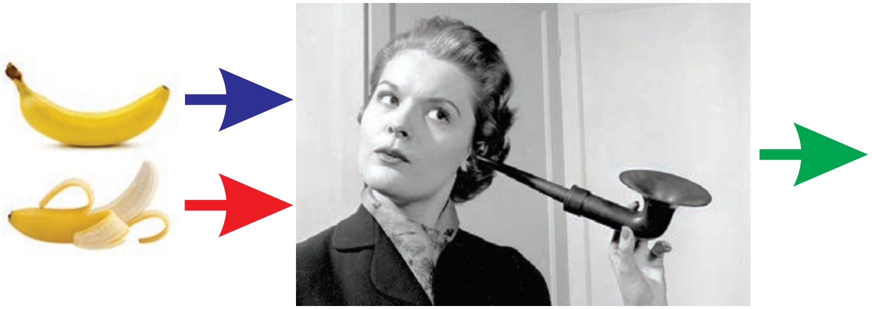
Testing should be method agnostic



1024



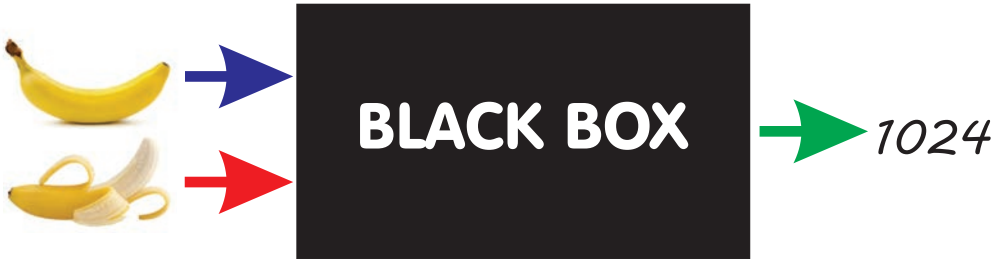
1,000,000

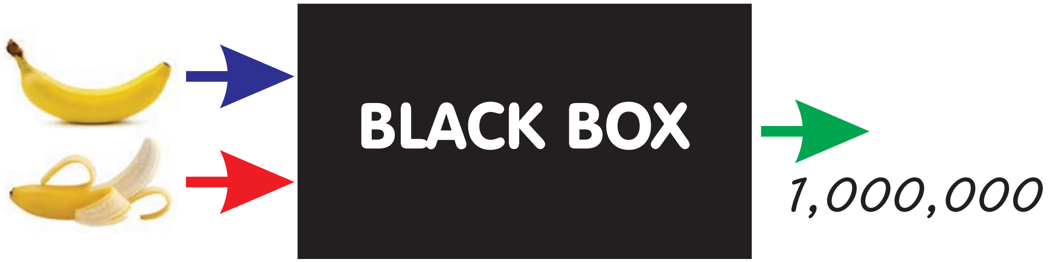


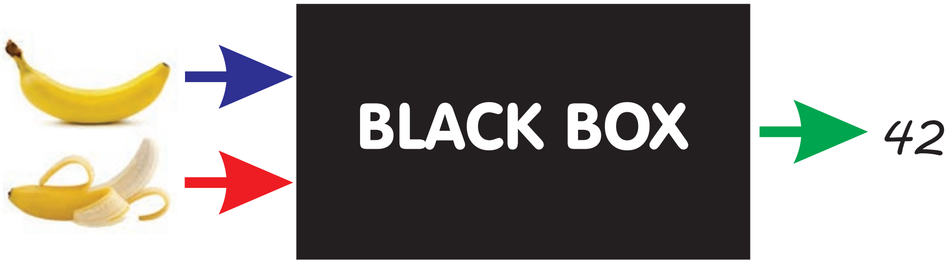
42

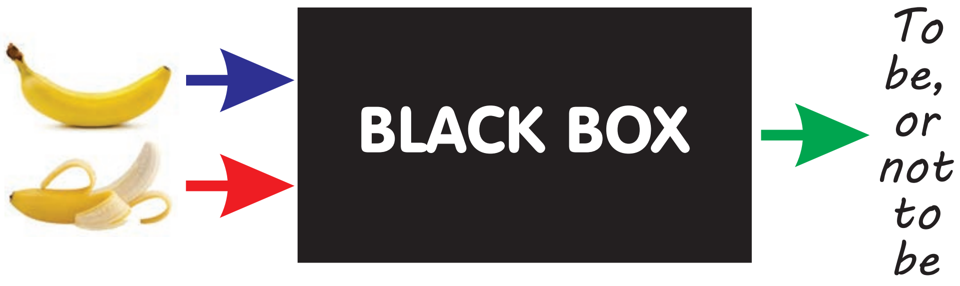
To
be,
or
not
to
be

Testing should be method agnostic



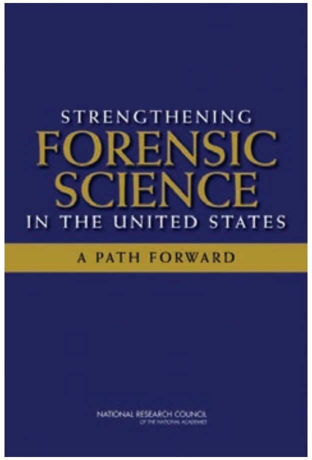






Validity and Reliability

- 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)

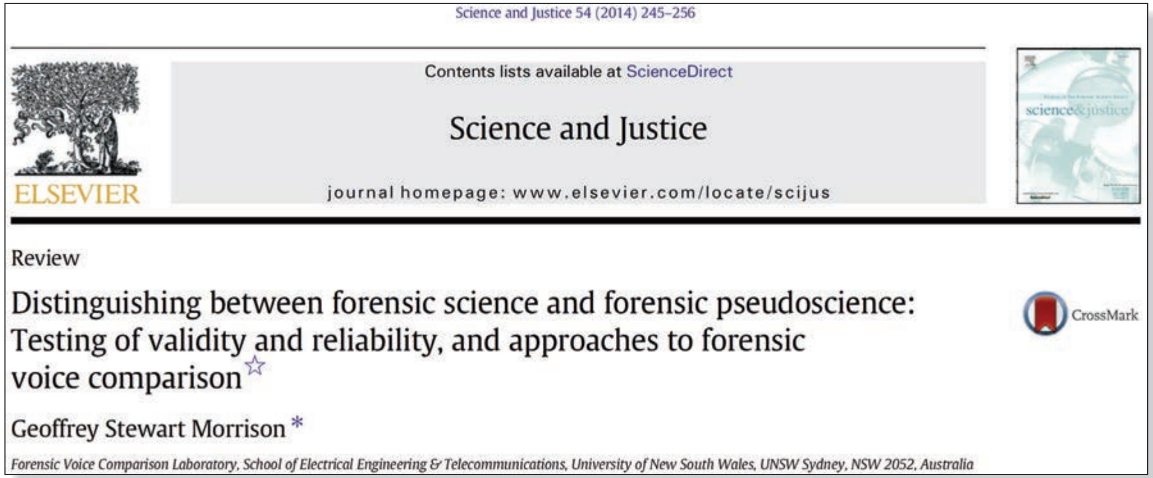
- “the conducting of validation studies of the performance of a forensic procedure” (p. 121)

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)

Forensic Science Regulator
Overseeing Quality

Validity and Reliability



Science and Justice 54 (2014) 245–256

Contents lists available at ScienceDirect

Science and Justice

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

ELSEVIER
CrossMark

Review

Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison☆

Geoffrey Stewart Morrison *

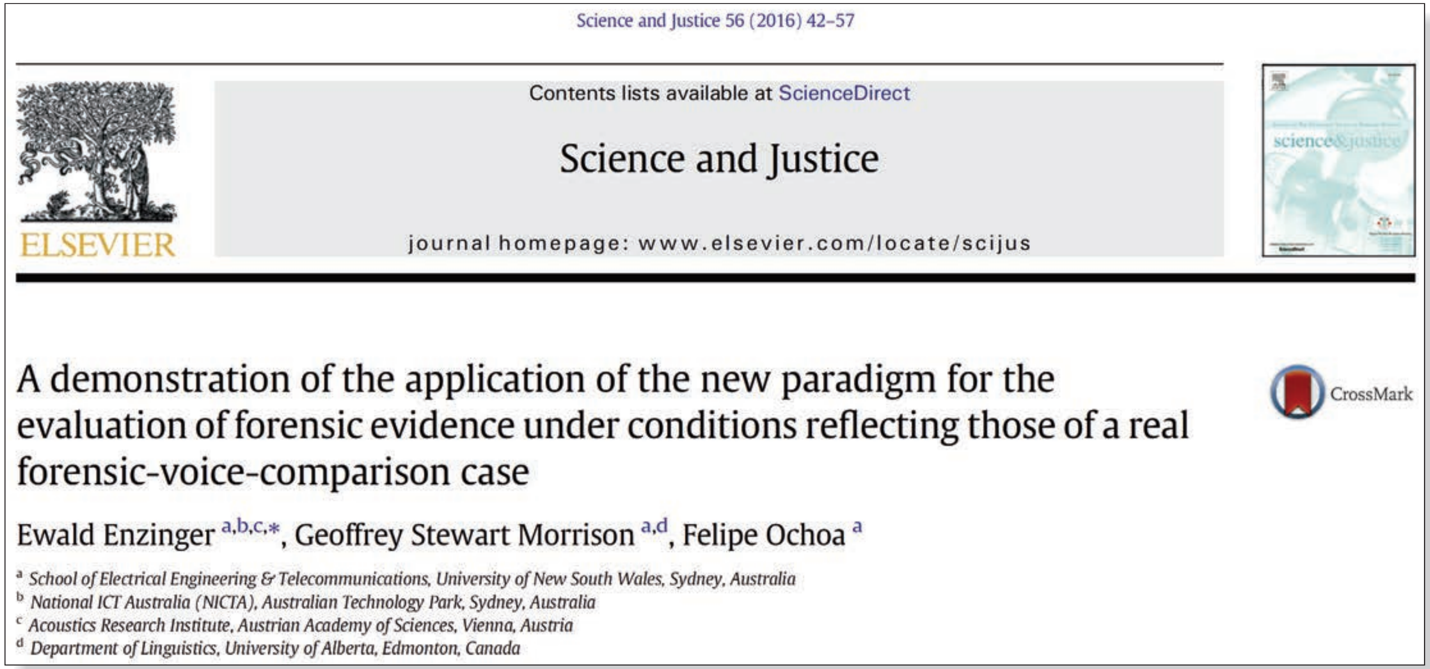
Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, UNSW Sydney, NSW 2052, Australia

Volume 56 Issue 2 March 2016
ELSEVIER
Journal of The Chartered Society of Forensic Sciences
science&justice
The Chartered Society of Forensic Sciences
Patron: The Hon. Mr Justice Sweeney
The Journal is also the official organ of the California Association of Criminalists
Available online at www.sciencedirect.com
ScienceDirect

Special issue on measuring and reporting the precision of forensic likelihood ratios

http://geoff-morrison.net/#special_issue_precision

An example of casework conducted within the new paradigm



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

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.

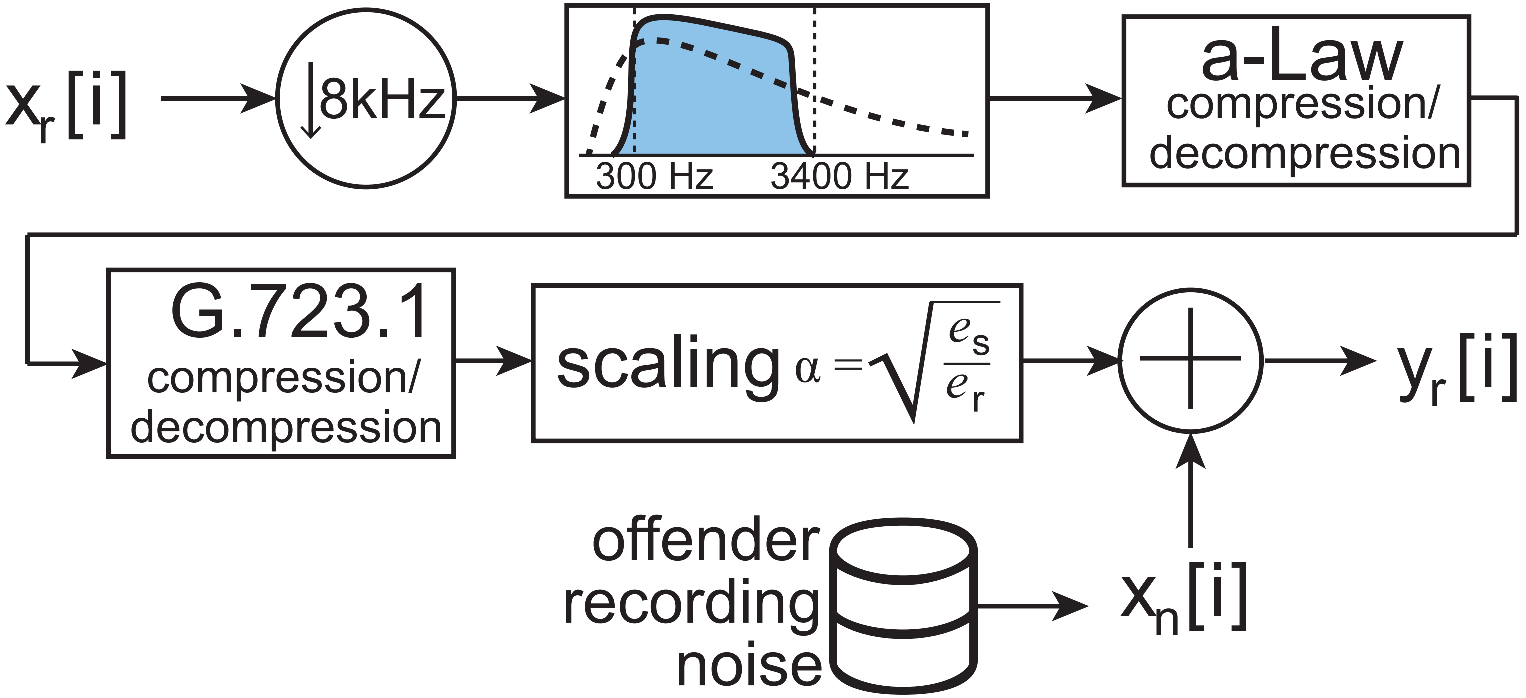
Defence hypothesis and relevant population adopted

- Relevant population chosen based on offender recording

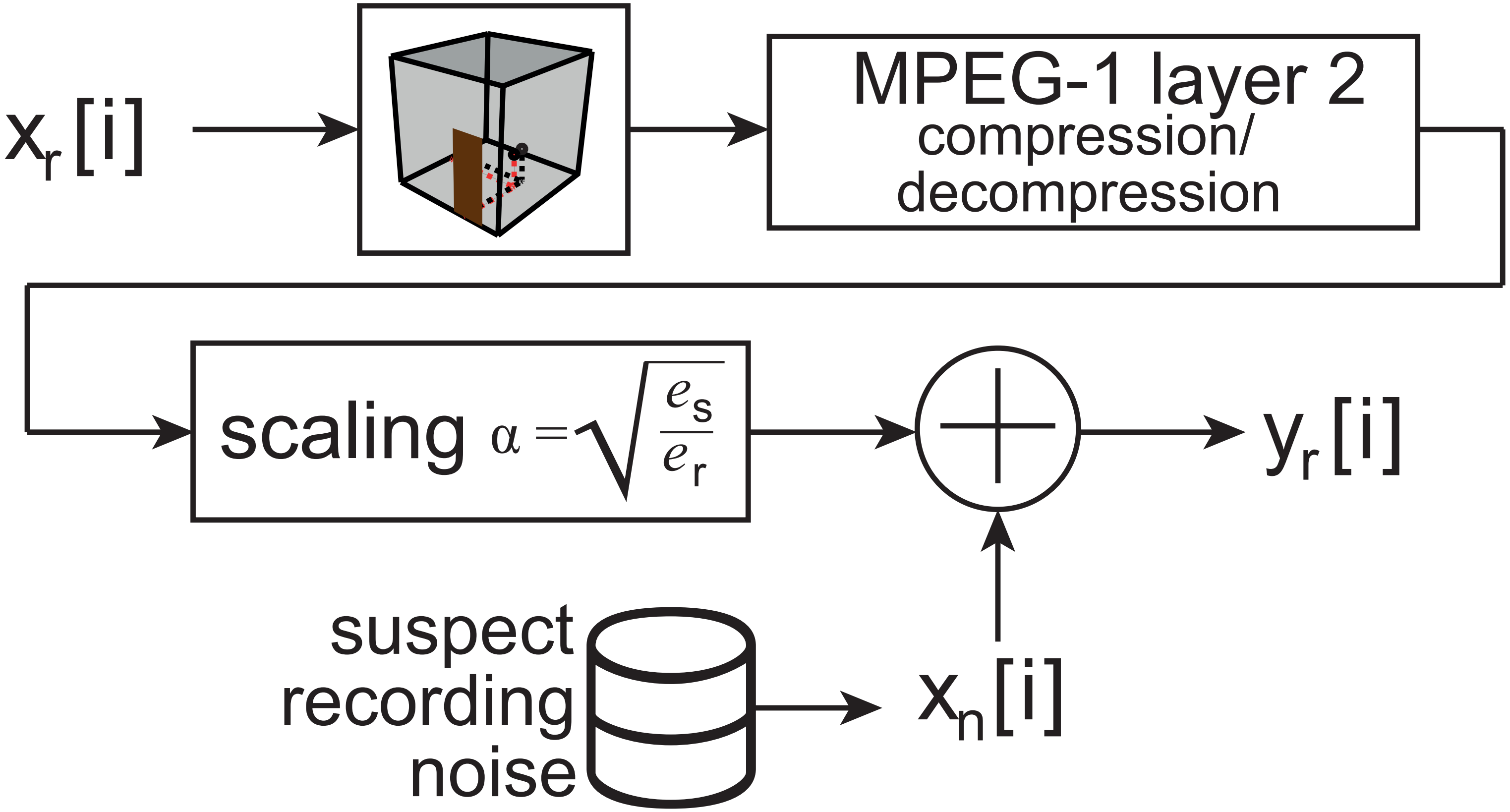
  Obvious that the speaker was
  - adult male
  - speaking Australian English

- 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


$x_r[i]$
↓8kHz
300 Hz
3400 Hz
a-Law compression/ decompression
G.723.1 compression/ decompression
scaling $\alpha = \sqrt{\frac{e_s}{e_r}}$
$y_r[i]$
offender recording noise
$x_n[i]$

Simulation of suspect-recording conditions



$x_r[i]$

MPEG-1 layer 2 compression/ decompression

scaling $\alpha = \sqrt{\frac{e_s}{e_r}}$

$y_r[i]$

suspect recording noise

$x_n[i]$

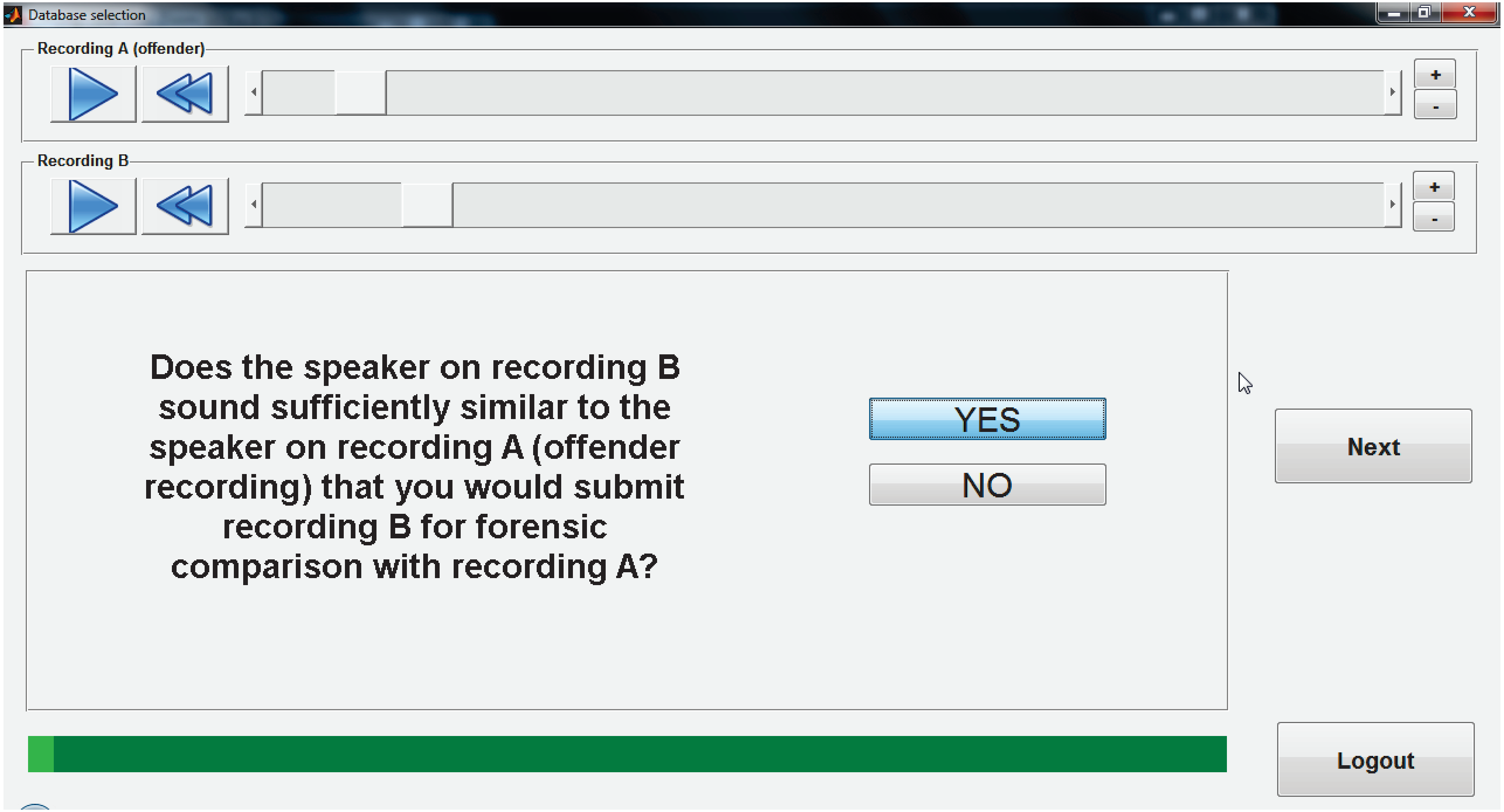
Selection of samples representative of the relevant population

- 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

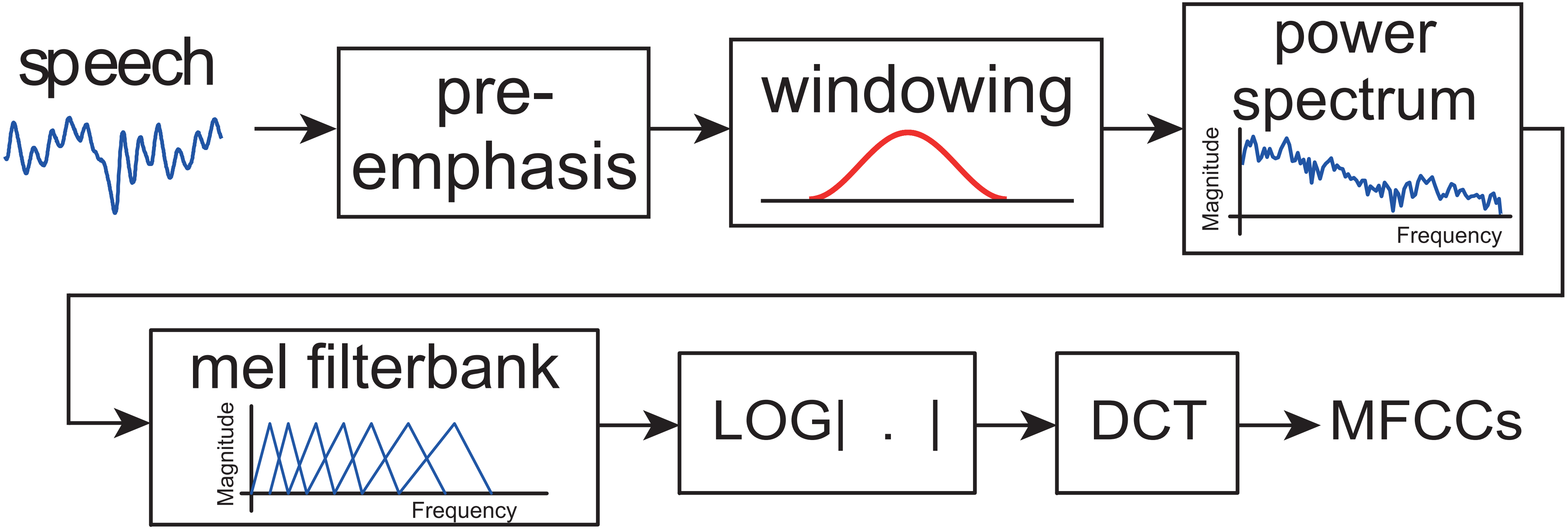
The number of speakers selected by *N* or more listeners.

number of listeners, *N*	11	10	9	8	7	6	5	4	**3**	2	1
number of speakers selected by *N* or more listeners	16	24	34	42	51	75	100	128	**166**	195	216

- 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


speech

pre-emphasis

windowing

power spectrum

mel filterbank

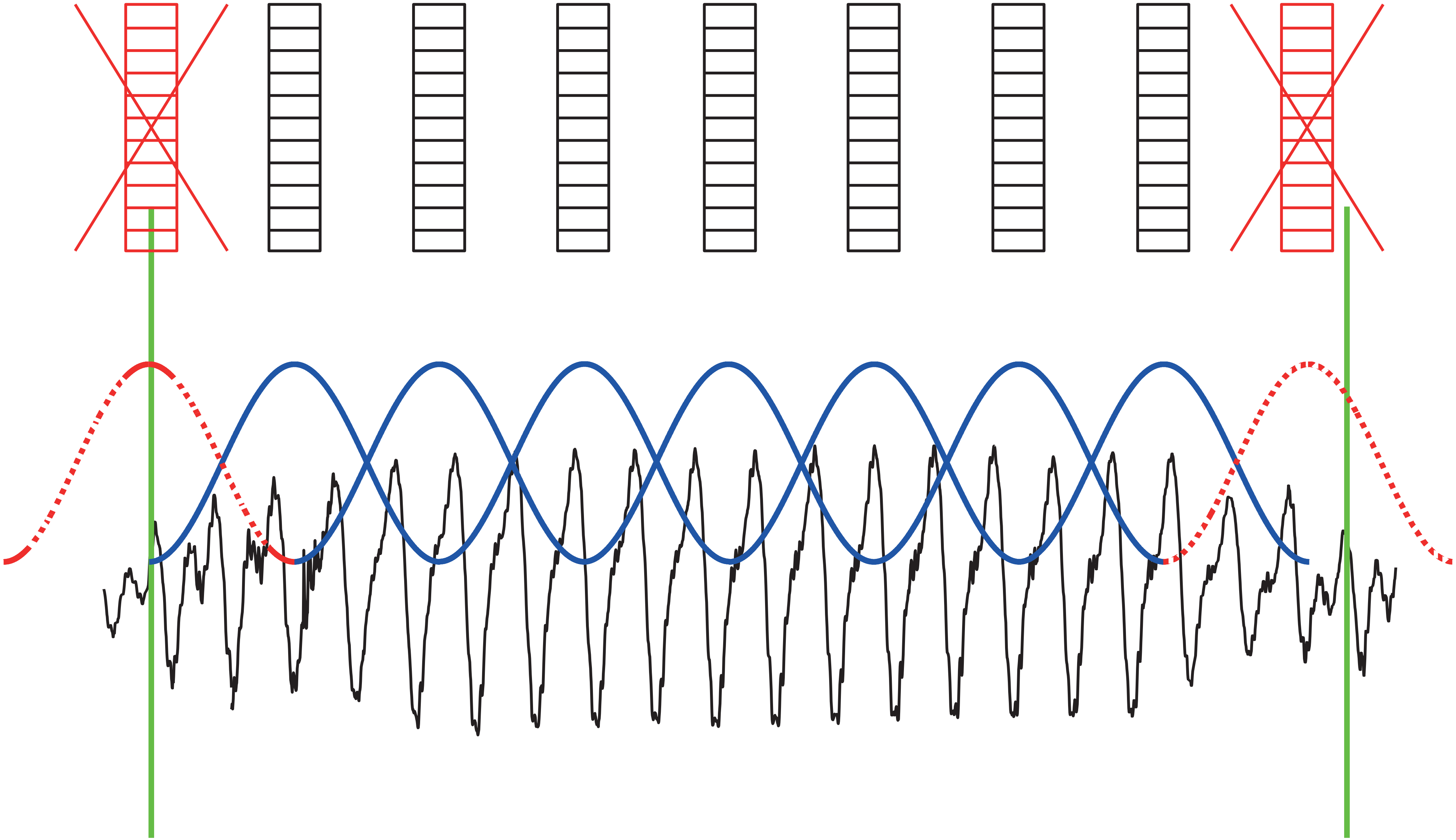
LOG| . |

DCT

MFCCs

– Suspect-condition and offender-conditions recordings made same durations as the actual suspect and offender recordings (in MFCC frames)

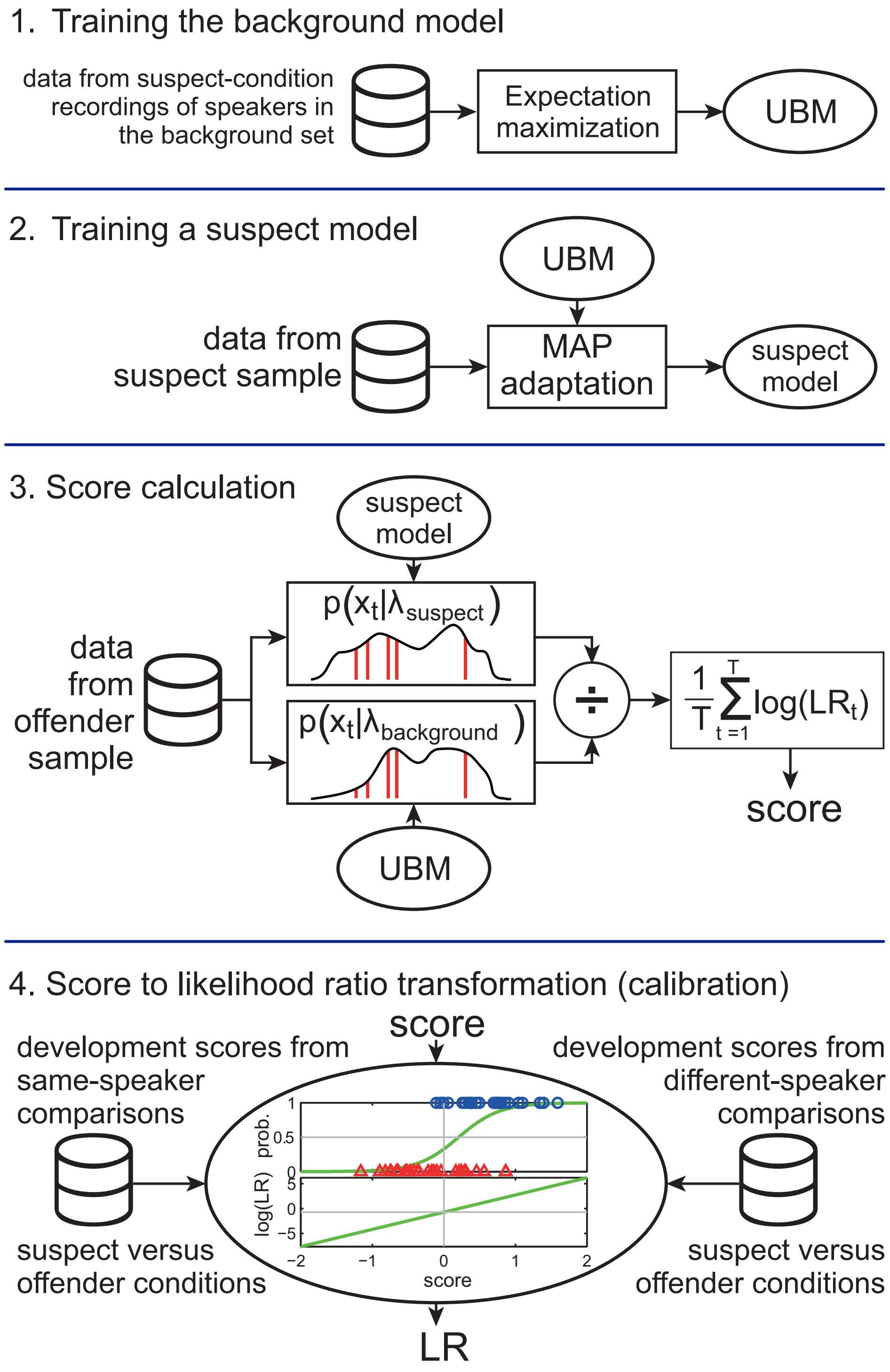
Quantitative acoustic measurements



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**
  - same-speaker pairs
  - different-speaker pairs



1. Training the background model

data from suspect-condition recordings of speakers in the background set

Expectation maximization

UBM

2. Training a suspect model

UBM

data from suspect sample

MAP adaptation

suspect model

3. Score calculation

suspect model

data from offender sample

$p(x_t|\lambda_{suspect})$

$p(x_t|\lambda_{background})$

UBM

$\div$

$\frac{1}{T}\sum_{t=1}^{T}\log(LR_t)$

score

4. Score to likelihood ratio transformation (calibration)

score

development scores from same-speaker comparisons

suspect versus offender conditions

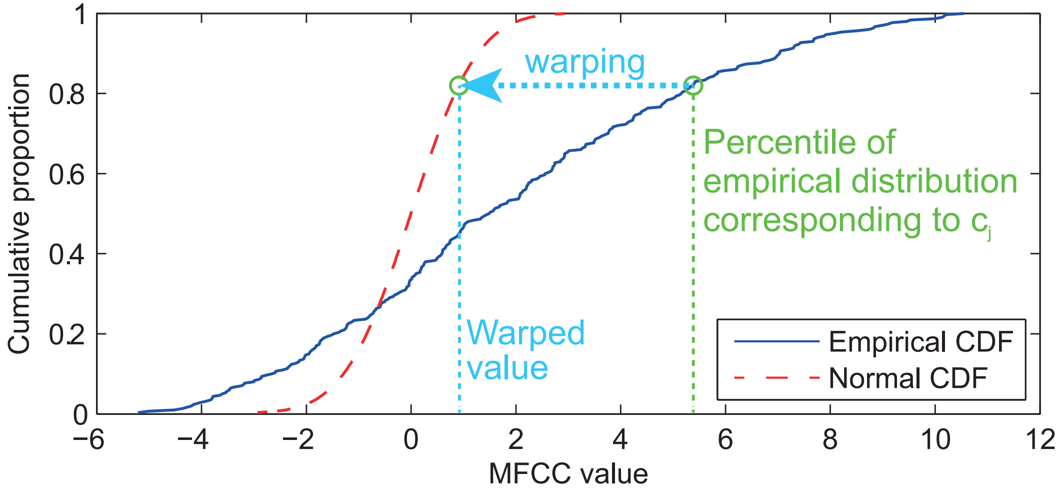
development scores from different-speaker comparisons

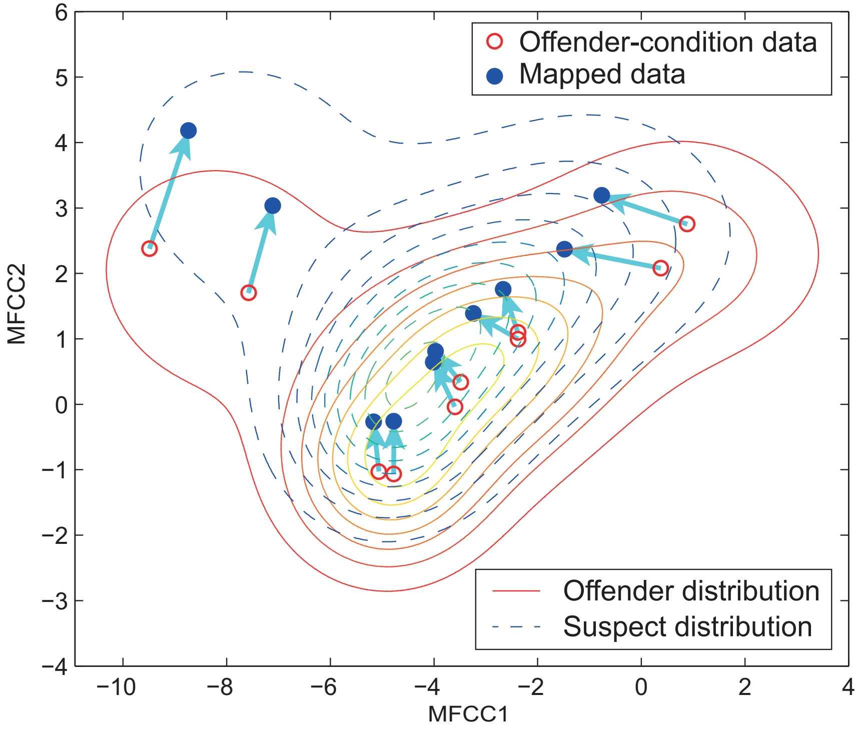
suspect versus offender conditions

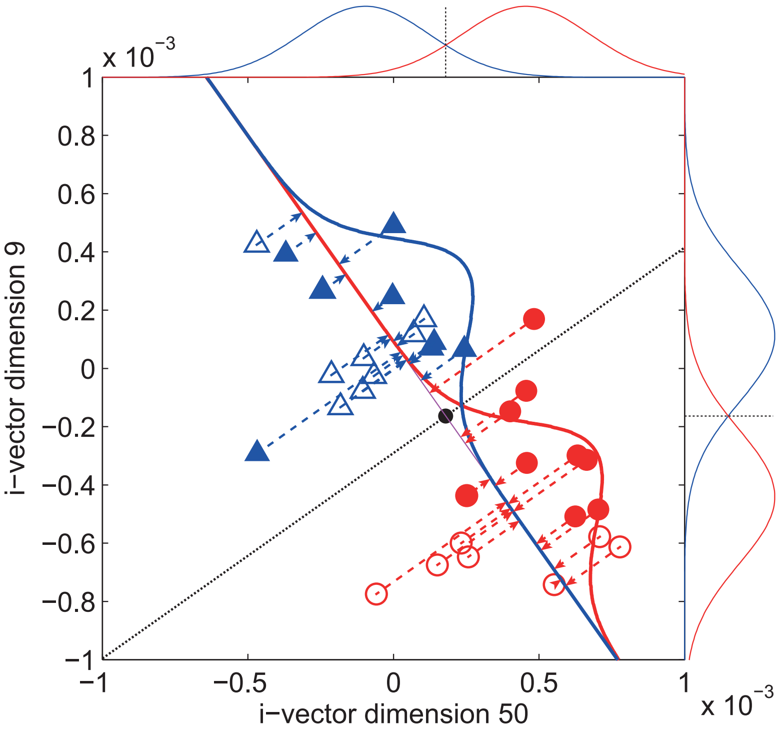
LR

Statistical models

- Mismatch compensation techniques
  - – feature warping
  - – probabilistic feature mapping
  - – canonical linear discriminant functions

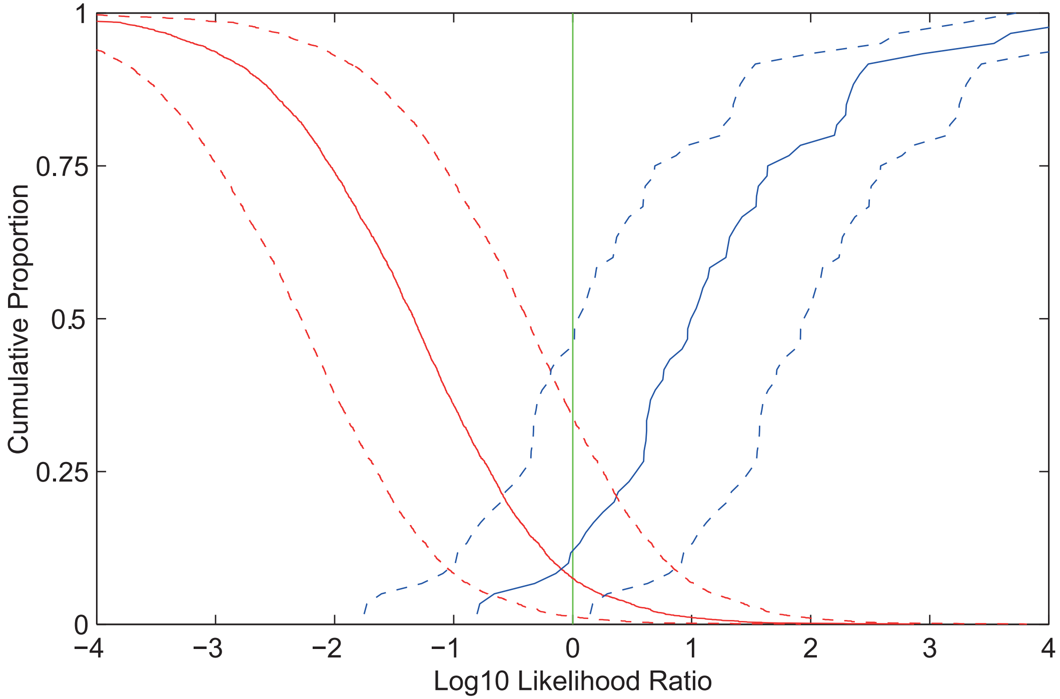






Test results

- Test data:
    - pairs of recordings from **sample of the relevant population**, one in **suspect condition**, the other in **offender condition**
    - same-speaker pairs
    - different-speaker pairs



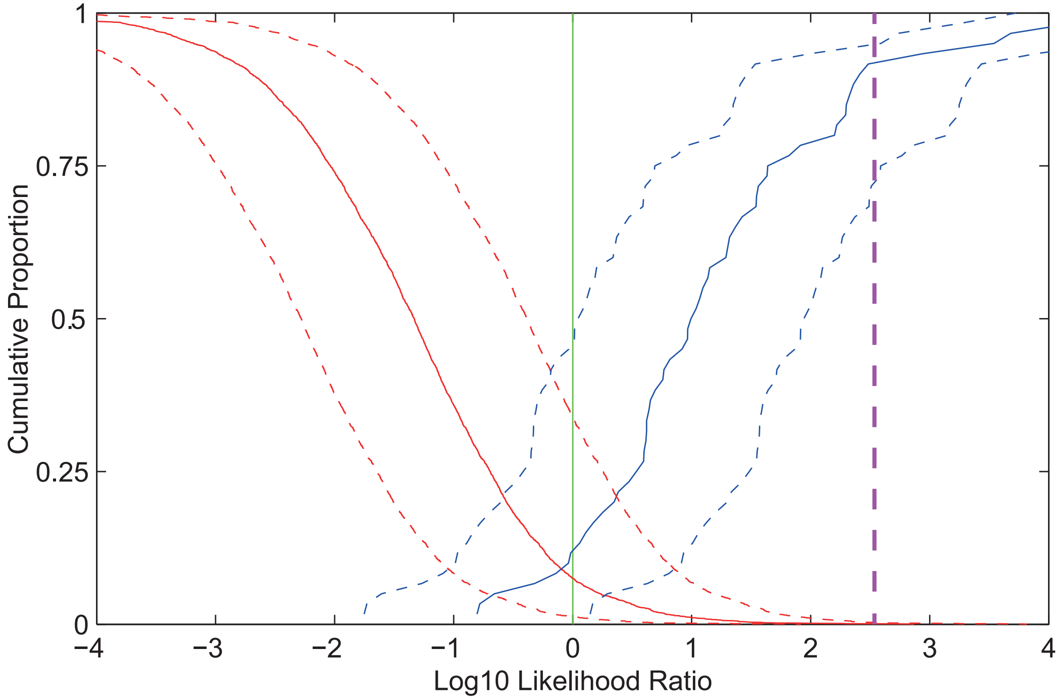
- C_{llr}-pooled: 0.423

- C_{llr}-mean: 0.344

- 95% CI: ±0.95

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



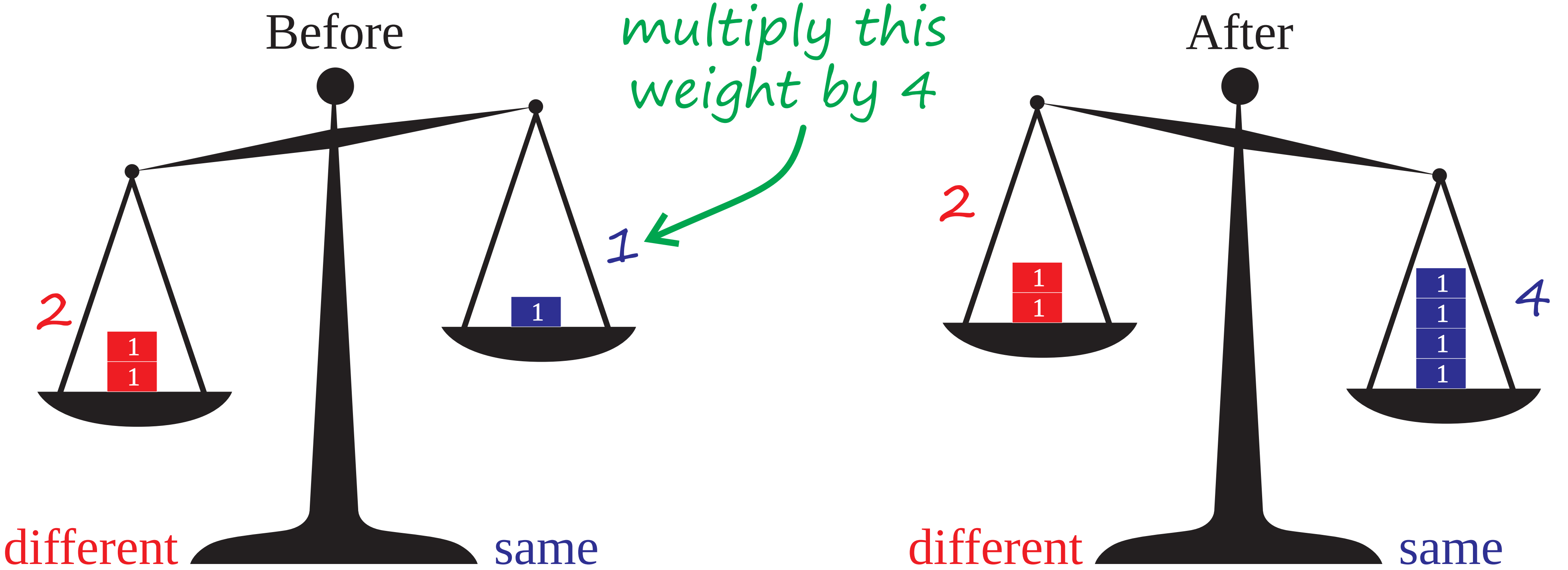
Conclusions

- Based on our calculations we estimate that the probability of obtaining the acoustic properties of the speech on the offender recording is **approximately 350 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 less than four in ten thousand (0.00033).

Example: The evidence is 4 time more likely given the same-speaker hypothesis than given the different-speaker hypothesis



if before you believed that the different-speaker hypothesis was 2 times more probable than the same-speaker hypotheses

now you believe that the same-speaker hypothesis is 2 times more probable than the different-speaker hypothesis

Thank You

http://forensic-voice-comparison.net/

http://forensic-evaluation.net/