Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case

Geoffrey Stewart Morrison
Ewald Enzinger

\[
p\left(\frac{E|H_p)}{p(E|H_d)}\right)
\]
Need for testing

- In forensic voice comparison, calls for validity and reliability to be empirically tested under casework conditions date back to the 1960s, but still go widely unheeded.

- Across all branches of forensic science, there is now increasing pressure to validate performance before analysis systems are used to assess strength of evidence for presentation in court
  - *Daubert v Merrell Dow Pharmaceuticals* [1993, 509 US 579]
  - National Research Council Report 2009
  - Forensic Science Regulator Codes of Practice 2014
  - ENFSI 2015 *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*
forensic_eval_01

- Open to operational forensic laboratories and research laboratories

- **Training and test data based on a real forensic case**
  - relevant population
  - speaking styles
  - recording conditions

- **Virtual Special Issue in *Speech Communication***
  - introductory paper includes rules
  - describe system and procedures in sufficient detail for replication
  - performance metrics and graphics
  - discussion and conclusion may include recommendations for practice
  - submissions accepted over a 2 year timeframe
- Casework conditions vary substantially from case to case

- *forensic_eval_01* evaluates systems under conditions reflecting those of one real case

- Results should not be assumed to be generalisable to other case conditions

- For each case, the validity and reliability of the system employed should be assessed under conditions reflecting those of that case
Forensic Voice Comparison 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
  
  - adult male Australian English speaker

- **Suspect recording**
  
  Police interview
  
  - reverberation
  
  - ventilation system noise
  
  - saved in a compressed format
Data

- Male Australian English speakers
- Multiple non-contemporaneous recordings per speaker
- Multiple speaking tasks per recording session
- High-quality audio

- **Offender condition**
  - information exchange task as input

- **Suspect condition**
  - interview task as input
Data

- **Training data:**
  - 423 recordings from 105 speakers
  - 191 recordings in offender condition
  - 232 in suspect condition

- **Test data:**
  - 223 recordings from 61 speakers
  - 61 recordings in offender condition
  - 162 in suspect condition
forensic_eval_01

- preliminary results from systems already tested on the forensic_eval_01 data
Enzinger & Morrison i-vector system

- 1st through 14th MFCCs + deltas
  – feature warping

- UBM
  – 512 Gaussians

- T-matrix
  – 400 or 200 dimensions

- i-vector domain mismatch compensation
  – canonical linear discriminant functions (aka LDA), 50 dimensions

- PLDA
  – full rank covariance for $\mathbf{B}$ and for $\mathbf{W}$

- score to likelihood ratio conversion (aka calibration)
  – logistic regression
Enzinger & Morrison i-vector system

- Generic data for training models which calculate scores
- Generic data for training mismatch compensation models in i-vector domain
- Case specific data for training score-to-LR model

- Case specific data for training models which calculate scores
- Case specific + generic data for training mismatch compensation models in i-vector domain
- Case specific data for training score-to-LR model
Enzinger & Morrison i-vector system

- **Generic data**
- **Case specific data**

95% credible interval (± order of magnitude)
Enzinger & Morrison i-vector system

- **Generic data**

- **Case specific data**
Batvox v4.1

- evaluated by David van der Vloed, Netherlands Forensic Institute

- **reference population data**
  - all 105 speakers (1 suspect-condition recording per speaker)
  - 30 selected by Batvox

- **imposter data**
  - none
  - all 105 speakers (1 offender-condition recording per speaker)
Batvox v4.1

The diagram shows a scatter plot with the following details:

- **C_{lir-mean}** vs. **95% credible interval (± order of magnitude)**
- Points are color-coded to represent different data sets:
  - ▲: all reference data + no imposter data
  - ▲: all reference data + imposter data
  - ▼: selected reference data + no imposter data
  - ▼: selected reference data + imposter data

The x-axis represents the 95% credible interval (± order of magnitude), while the y-axis represents the C_{lir-mean} values.
Batvox v4.1

95% credible interval (± order of magnitude)

- 30 reference speakers
- 105 reference speakers

Legend:
- △ all reference data + no imposter data
- ▲ all reference data + imposter data
- ▽ selected reference data + no imposter data
- ▼ selected reference data + imposter data
Batvox v4.1

95% credible interval (± order of magnitude)

- all reference data + no imposter data
- all reference data + imposter data
- selected reference data + no imposter data
- selected reference data + imposter data

no imposters
105 imposters
Batvox v4.1

105 reference speakers
105 imposters

95% credible interval (± order of magnitude)
Batvox v4.1

105 reference speakers

105 imposters

30 reference speakers
Eskerrrik Asko

http://geoff-morrison.net/
http://forensic-evaluation.net/
Best of

- Batvox v4.1
- Enzinger & Morrison
Best of

Batvox v4.1

Enzinger & Morrison