A demonstration of the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case

Ewald Enzinger
Geoffrey Stewart Morrison

Acknowledgement of Funding

- Data collection supported by the Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142. Unless otherwise explicitly attributed, the opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations.

Introduction

- Demonstration of the evaluation of forensic evidence under conditions reflecting those of an actual forensic-voice-comparison case

- Casework recordings are not being used
  - Speaker selected from a database as mock offender/suspect
  - Use circumstances of the case

- Practical illustration of FVC performed under the following paradigm…
  (see Morrison, 2014; Morrison & Stoel, 2014)

Paradigm for evaluation of FVC evidence

- Likelihood-ratio framework
  - Statement of strength of the evidence as an answer to a specific question $\mathrm{LR} = \dfrac{p(E|H_p)}{p(E|H_d)}$
- Use of data representative of the relevant hypotheses/populations, quantitative measurements, and statistical models
- Testing of validity and reliability under conditions reflecting those of the case

Case background

- Offender recording:
  - landline telephone call made to a call centre

- Suspect recording:
  - Questioning by the police at a police station

Methodology for practical implementation

- Definition of relevant hypotheses and populations
- Collection / simulation of data relevant to the case
- Development of FVC system to calculate a likelihood ratio given the hypotheses
- Empirical testing of validity and reliability under recording conditions reflecting those of the case
- Evaluation of the likelihood ratio for the actual suspect and offender samples

Defining the relevant hypotheses

- What is the likelihood of getting the measured acoustic properties of the voice on the offender recording if the speaker on that recording were the suspect? (*probability of evidence given prosecution hypothesis*)

  vs.

- What is the likelihood of getting the measured acoustic properties of the voice on the offender recording if the speaker on that recording were not the suspect but some other speaker from the relevant population? (*probability of evidence given defence hypothesis*)

We need a sample of voice recordings of people from the relevant population...

- to assess the typicality of the acoustic properties of the offender recording with respect to the defence hypothesis
- to develop a forensic-voice-comparison system
  - recording-condition mismatch compensation
  - score to likelihood ratio transformation (calibration)
- to empirically test its validity and reliability given the hypotheses under recording conditions reflecting those of the case

Selection of data reflecting the hypotheses

- No dispute that suspect and speaker on offender recording were adult male Australian English speakers
- In the case a police officer listened to suspect/offender samples and decided that the voice on the suspect recording sounded sufficiently similar to the voice on the offender recording that it was worth submitting them for forensic comparison with each other

➔ Sampling from database of male Australian English voice recordings by a panel of human listeners

Initial database

- 230 male Australian English speakers
- Multiple non-contemporaneous recordings
- Speaking style
  - Offender: Information exchange task over telephone
  - Suspect: Pseudo-police-style interview
- High-quality recording conditions
  - Simulation of suspect/offender sample recording conditions (reverberation, noise, compression, etc.)

Simulation of recording conditions

- Offender: landline telephone call made to a call centre
  - Telephone transmission, background noise, compression
- Simulation:
  - Resampling at 8 kHz
  - Telephone bandpass filter (ITU-T rec G.151)
  - a-Law companding algorithm (ITU-T rec G.711)
  - Compression using G.723.1 codec
  - Adding noise taken from non-speech sections of offender recording

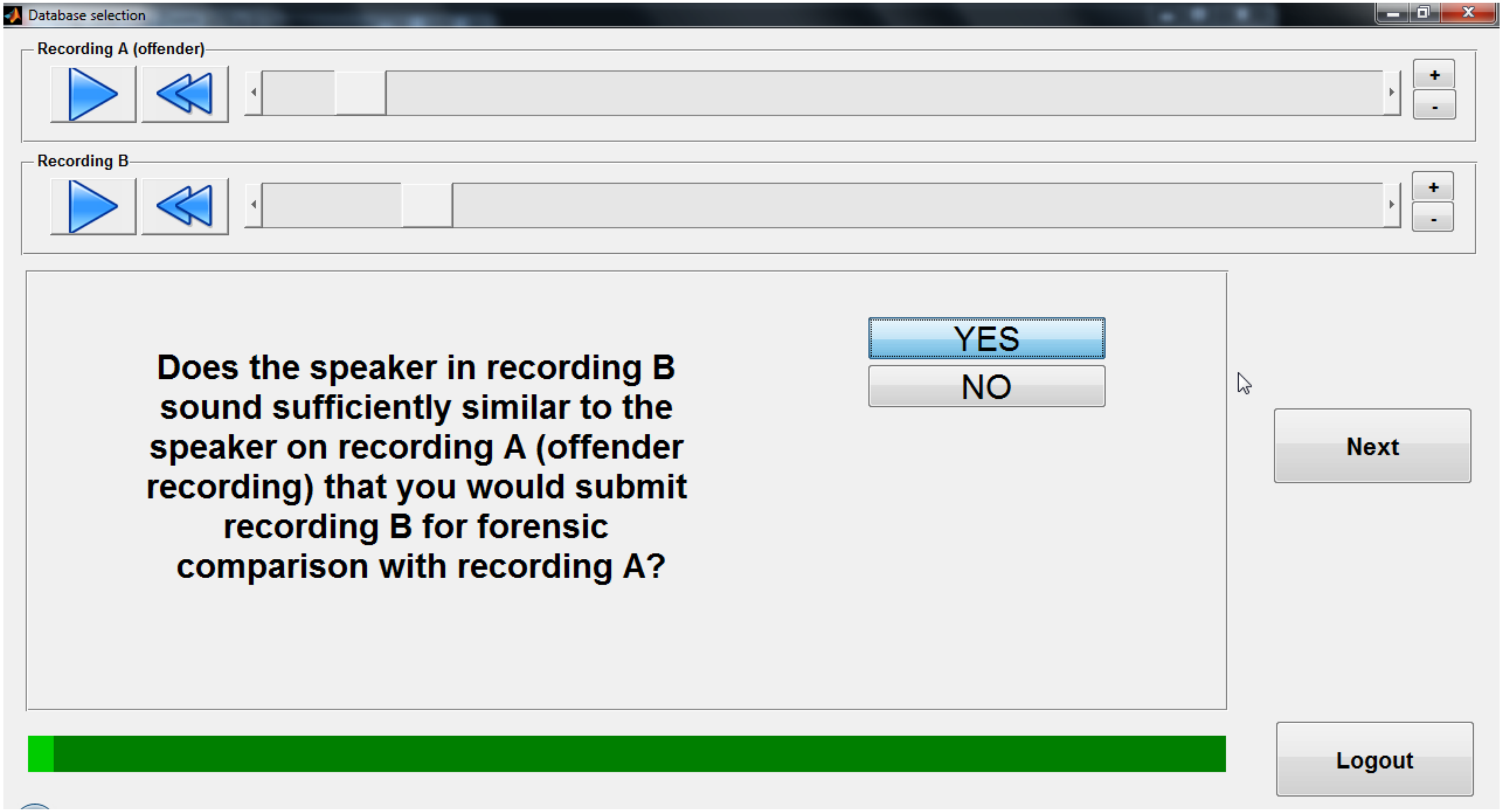
Simulation of recording conditions

- Suspect: recording of police interview
  - reverberant room, ventilation noise, compression
- Simulation:
  - Reverberation simulation using information from location where police interview had been conducted (room dimensions, location of microphone/speaker)
  - Compression using MPEG-1 layer 2 codec
  - Adding noise taken from non-speech sections of suspect recording

Selection of data reflecting the hypotheses

- Recruited 11 human listeners
  - First (and only) language was Australian English
  - Born and raised in Australia
- Listen to the offender sample as well as suspect-condition samples from the database

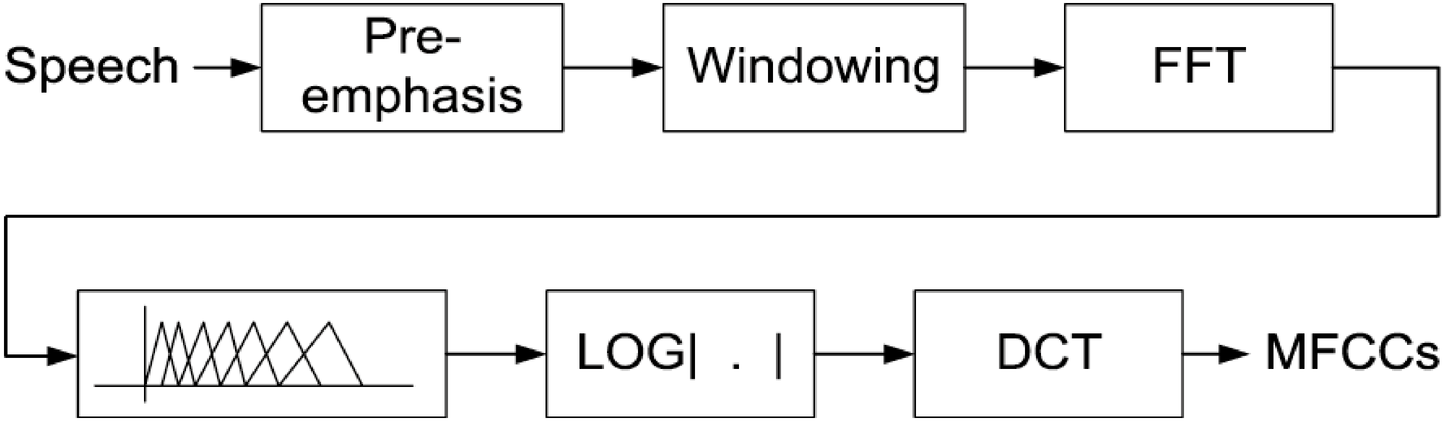
Selection of data reflecting the hypotheses



Does the speaker in 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?

Forensic-voice-comparison system

- Acoustic analysis:
  - Mel frequency cepstral coefficients (MFCCs)
  - 1st -14th coefficients every 10 ms over the speech-active portion
  - Delta coefficients



Forensic-voice-comparison system

- Statistical analysis:
    - Gaussian mixture model – Universal background model (GMM-UBM)
    - 512 Gaussian mixture components
    - Trained using data from suspect-condition recordings of 44 background speakers

Forensic-voice-comparison system

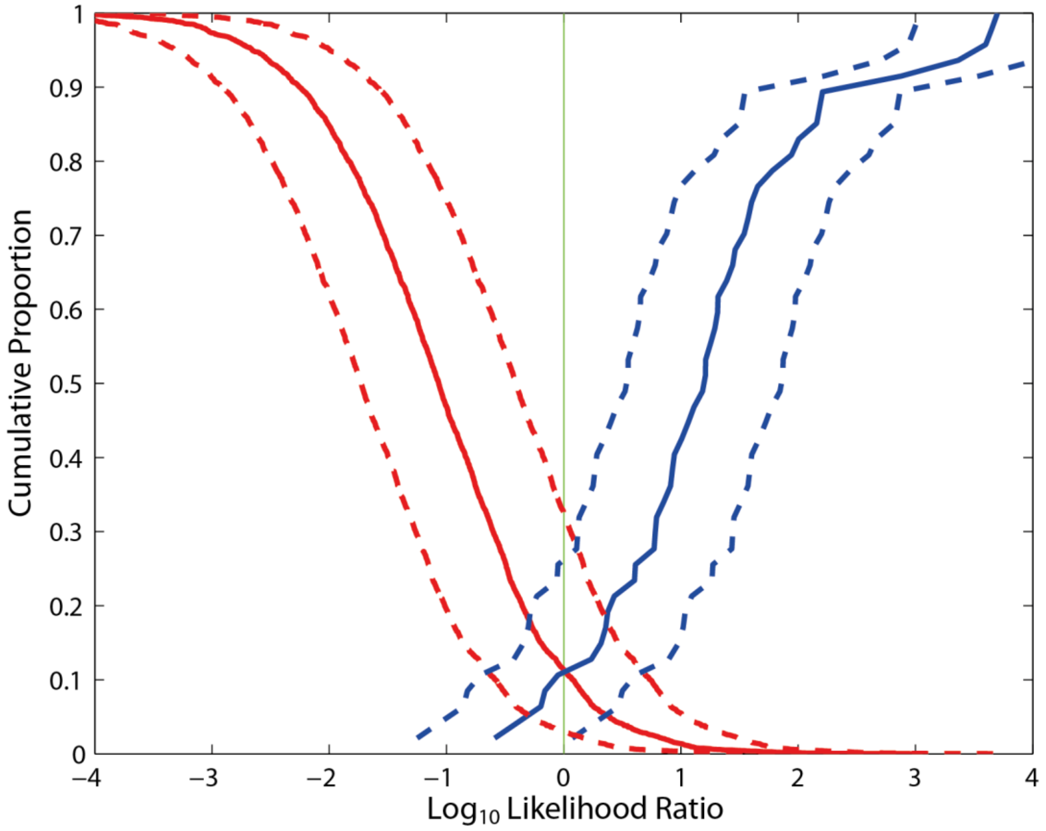
- Recording-condition mismatch compensation
  - see poster session II, tomorrow at 2 p.m.

    "Mismatch compensation in the evaluation of evidence under conditions reflecting those of an actual forensic-voice-comparison case"
- Score to likelihood ratio transformation
  - Logistic regression calibration
  - Trained from scores of comparisons between 61 development speakers

Testing of validity and reliability

- Tests using 60 held-out test speakers
  - after freezing the FVC system
  - before calculating the likelihood ratio in the case
- Tippett plots
- Validity (C_{llr}) and reliability (credible interval)

Testing of validity and reliability



$C_{llr} = 0.347$

95% CI = 0.66

Conclusion

- Illustration of a methodology for implementation of a paradigm for the evaluation of forensic evidence, under conditions reflecting those of an actual case
- Intended to spur discussion on how to implement these principles in practice
- Methodology can be adapted to other FVC casework and using other kinds of acoustic measurements or statistical models

Thank you

References

Morrison, G. S., & Stoel, R. D. (2014). 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? Aus J Forensic Sci 46, 282–292. doi:10.1080/00450618.2013.833648

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. Sci. Justice 54, 245–256. doi:10.1016/j.scijus.2013.07.004