Implementation of the new paradigm in forensic-voice-comparison practice

Ewald Enzinger

Problems with forensic-voice-comparison research and practice

- Current practice in forensic comparison sciences lacks testing of validity and reliability (NRC report 2009, NIST/NIJ report 2011)

- Need for empirical testing under conditions reflecting those of forensic casework recognised since the 1960s (Bolt et al. 1970)

Paradigm for Forensic Voice Comparison

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

Aims

- Establish methodology for performing practical forensic voice comparison in this paradigm
- Develop mismatch compensation techniques
  - Adapt methods from automatic speaker recognition
- Demonstration based on two case studies with conditions taken from real forensic-voice-comparison cases

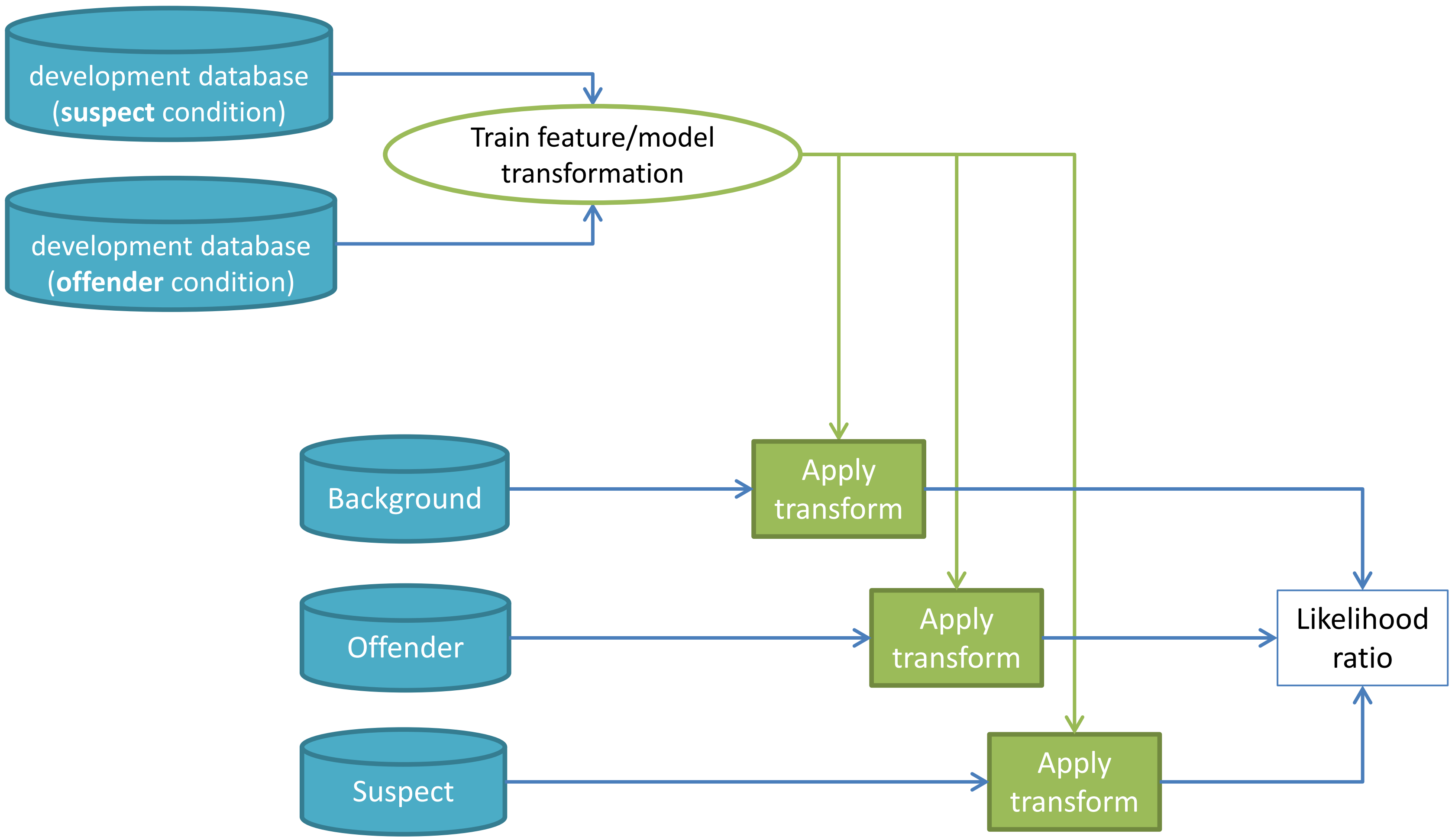
Methodology

1. Ascertain the competing hypotheses
2. Simulate suspect and offender conditions
3. Obtain sample of relevant population
4. Develop forensic-voice-comparison system
   - Feature extraction, mismatch compensation, statistical modelling
5. Test validity and reliability on held-out data
6. Calculate LR for offender and suspect sample

Mismatched conditions

- Samples collected under different conditions
- Known to severely degrade performance
- Application of mismatch compensation in FVC:
  + Knowledge of suspect and offender conditions
    - Transmission channel (GSM vs. landline)
    - Level of background noise, reverberation
  – Limited amount of data
    - Selection of appropriate database for a given case

Mismatch compensation in FVC



Case studies

- Demonstration of methodology based on two case studies using conditions taken from real forensic-voice-comparison cases
  - Case 1: Mismatched distances
  - Case 2: Australian English males

Case study 1: Mismatched distances

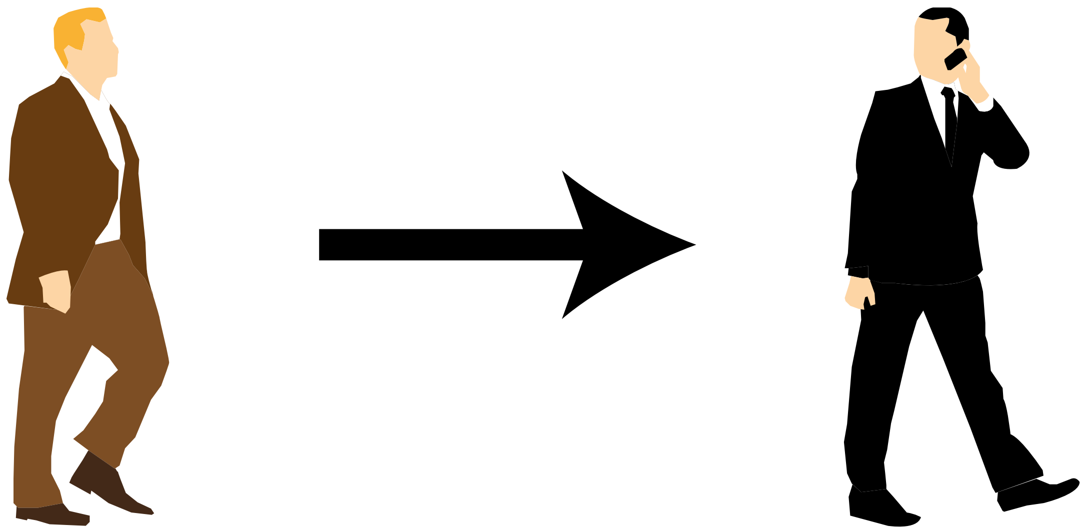
- 2 speakers, 1 speaking on a mobile phone

Speaker A

Speaker B

Case study 1: Mismatched distances

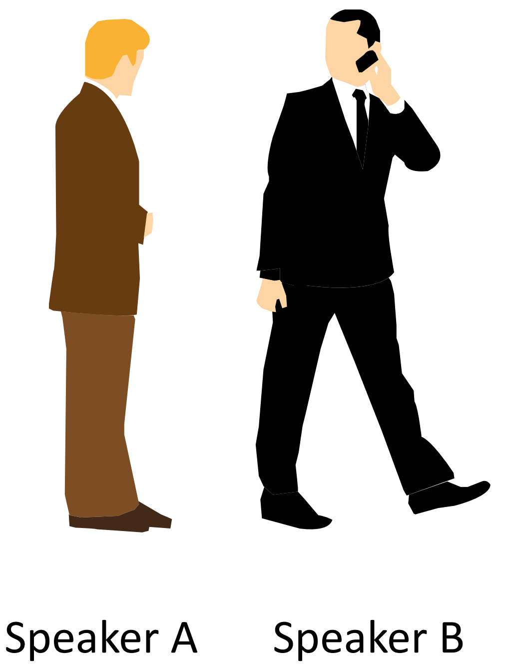
- Later, speaker B moves closer to the telephone



Speaker A　　　　Speaker B

Case study 1: Mismatched distances

- Shortly after: Section where identity of the speaker is unknown



Case study 1: Mismatched distances

- Hypotheses considered:
  - the questioned utterance was spoken by speaker A

    vs.
  - the questioned utterance was spoken by speaker B
- Near-far mismatch between training data for speaker A and speaker B
  - Speaker A: 90% of undisputed speech (27 s) is **far**, 10% **near**
  - Speaker B: all of undisputed speech (47 s) is **near**
  - Questioned utterance: (2.08 s) is **near**

Case study 1: Mismatched distances

Completed work:

- Simulation of near/far recording conditions
- Developed forensic-voice-comparison system
- Procedure for testing validity and reliability
- Testing the effects of distance mismatch
- Apply mismatch compensation
  - Likelihood-ratio domain bias compensation
  - Feature mapping (mean feature vector offset)
  - Linear discriminant function analysis transform

Case study 1: Mismatched distances

- Portions of preliminary work published in:
  Enzinger, E. (2013). Mismatched distances from speakers to telephone in a forensic-voice-comparison case. Proceedings of the 21st International Congress on Acoustics (ICA), June 2–7, Montréal, Canada (POMA Volume 19, pp. 060039). doi:10.1121/1.4805425.
- Journal paper to be submitted by the end of 2013

Mismatched distances from speakers to telephone in a forensic-voice-comparison case☆

Ewald Enzinger[a,b,c], Geoffrey Stewart Morrison[a]

[a] *School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia*
[b] *National ICT Australia (NICTA), Australian Technology Park, Sydney, NSW 1430, Australia*
[c] *Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, A-1040 Vienna, Austria*

Abstract

In a forensic-voice-comparison (FVC) case, one speaker (A) was standing a short distance away, and another (B) was talking on a mobile telephone. Later, A moved closer to the telephone. Shortly thereafter, there was a section of speech where the identity of the speaker was disputed. All material for training an FVC-system could be extracted from this single recording, but there was a near-far mismatch: Training data for A were mostly far, training data for B were near, and the disputed speech was near. Based on the conditions of this FVC case we demonstrate a methodology for handling forensic casework in the new paradigm for the evaluation of forensic evidence, using the likelihood ratio framework, quantitative measurements, and statistical models. Using a set of development speakers we investigate the effect of mismatched distances to the microphone and demonstrate three methods for compensation. Finally a procedure is described for addressing the degree of validity and reliability of an FVC system under such conditions, prior to it being applied to the section of questioned identity.

Keywords: Forensic voice comparison, likelihood ratios, validity, reliability, distance mismatch, mismatch compensation

Case 2: Australian English males

- Offender sample: telephone recording
- Suspect sample: police interview
- Hypotheses considered:
  - 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)

Case 2: Australian English males

Completed work:

- Processing of Australian English database
- Simulation of suspect and offender conditions
- Developed forensic-voice-comparison system
- Apply mismatch compensation
  - Feature warping (Pelecanos & Sridharan, 2001)
  - Probabilistic feature mapping (Reynolds, 2003; Mak et al., 2007)
  - Feature-domain nuisance attribute projection (Campbell et al., 2008)
  - Feature-domain latent factor analysis (Vair et al., 2006; Castaldo et al., 2007; Campbell et al., 2008)

Case 2: Australian English males

Work to be done:

- Sampling of the relevant population:

  "Speakers who sound sufficiently similar to the voice of questioned identity that a police officer thinks that recordings of these speakers also sound sufficiently similar to the voice of questioned identity that it worth submitting them for forensic analysis."

  - HREA ethics application ready for submission
  - Contacts at Victoria Police and Queensland Police Service facilitate recruitment

Publications since last review

Journal articles:

Zhang, C., Morrison, G. S., **Enzinger, E.**, & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. Speech Communication, 55 (6), 796–813. doi:10.1016/j.specom.2013.01.011.

Zhang, C., Morrison, G. S., Ochoa, F., & **Enzinger, E.** (2013). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. Journal of the Acoustical Society of America, 133 (1), EL54–EL60. doi:10.1121/1.4773223.

Conference proceedings (reviewed by paper):

Enzinger, E. (2013 in press). Testing the validity and reliability of forensic voice comparison based on reassigned time-frequency representations of Chinese /iau/. Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 18–21 November, Guangzhou, China.

Enzinger, E. & Kasess, C. H. (2013). Experiments on using Vocal Tract Estimates of Nasal Stops for Speaker Verification. Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), Cluj-Napoca, Romania.

Enzinger, E. & Morrison, G. S. (2012). The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012), 3–6 December, Sydney, Australia, pp. 137–140.

Conference proceedings (reviewed by abstract):

Enzinger, E. (2013). Mismatched distances from speakers to telephone in a forensic-voice-comparison case. Proceedings of the 21st International Congress on Acoustics (ICA), June 2–7, Montréal, Canada (POMA Volume 19, pp. 060039). doi:10.1121/1.4805425.

Zhang, C. & **Enzinger, E.** (2013). Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/. Proceedings of the 21st International Congress on Acoustics (ICA), June 2–7, Montréal, Canada (POMA Volume 19, pp. 060044). doi:10.1121/1.4798793.

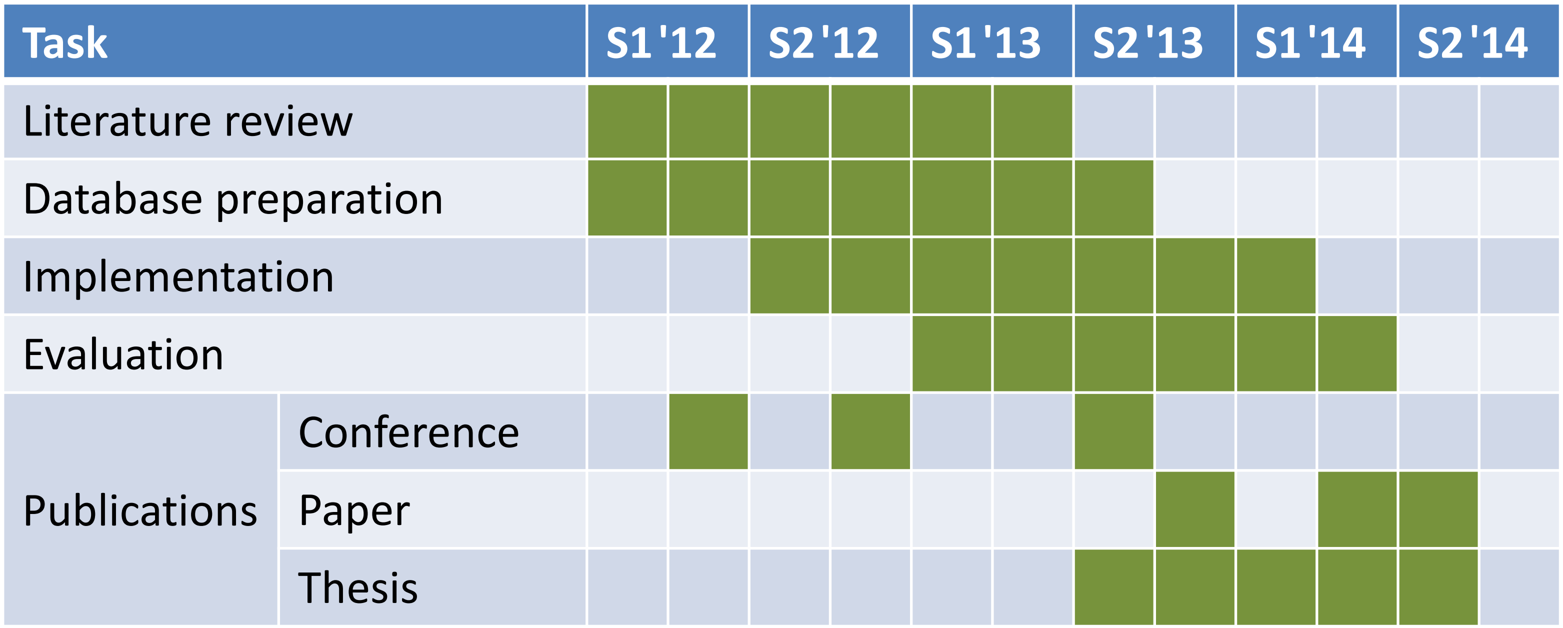
Invited conference proceedings:

Grigoras, C., Smith, J. M., Morrison, G. S., & **Enzinger, E.** (2013). Forensic audio analysis – Review: 2010–2013. In: NicDaéid, N. (Ed.), Proceedings of the 17th International Forensic Science Managers' Symposium, Lyon (pp. 612–637). Lyon, France: Interpol.

Goals for the next 12 months

- Finish analysis of Case 2 with results from database selection experiment
  - All software/procedures are already implemented
  - Just need to plug-in results and re-run analysis
- Submit journal paper on Case 1 by end of Dec 2013
- Submit journal paper on Case 2 by end of June 2014
- Send drafts of individual sections of the thesis to supervisor (schedule to be proposed)
- Thesis ready for submission by end of June 2014 *

Project plan



References

National Research Council (2009). Strengthening Forensic Science in the United States: A Path Forward. National Academies Press, Washington, DC. http://www.nap.edu/catalog.php?record_id=12589.

Expert Working Group on Human Factors in Latent Print Analysis (2012). Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach. US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD. http://www.nist.gov/manuscript-publication-search.cfm?pub_id=910745.

Bolt, R. H., Cooper, F. S., David Jr., E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1970). Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes. Journal of the Acoustical Society of America 47, 597-612.

Pelecanos, J. & Sridharan, S. (2001). Feature warping for robust speaker verification. In: Proc. Odyssey, Crete, Greece, pp. 213–218.

Reynolds, D. A. (2003). Channel robust speaker verification via feature mapping. In: Proc. ICASSP, Vol. 2, Hong Kong, pp. 53–56.

Mak, M.-W., Yiu, K.-K., & Kung, S.-Y. (2007). Probabilistic feature-based transformation for speaker verification over telephone networks. Neurocomputing 71 (1-3), 137–146.

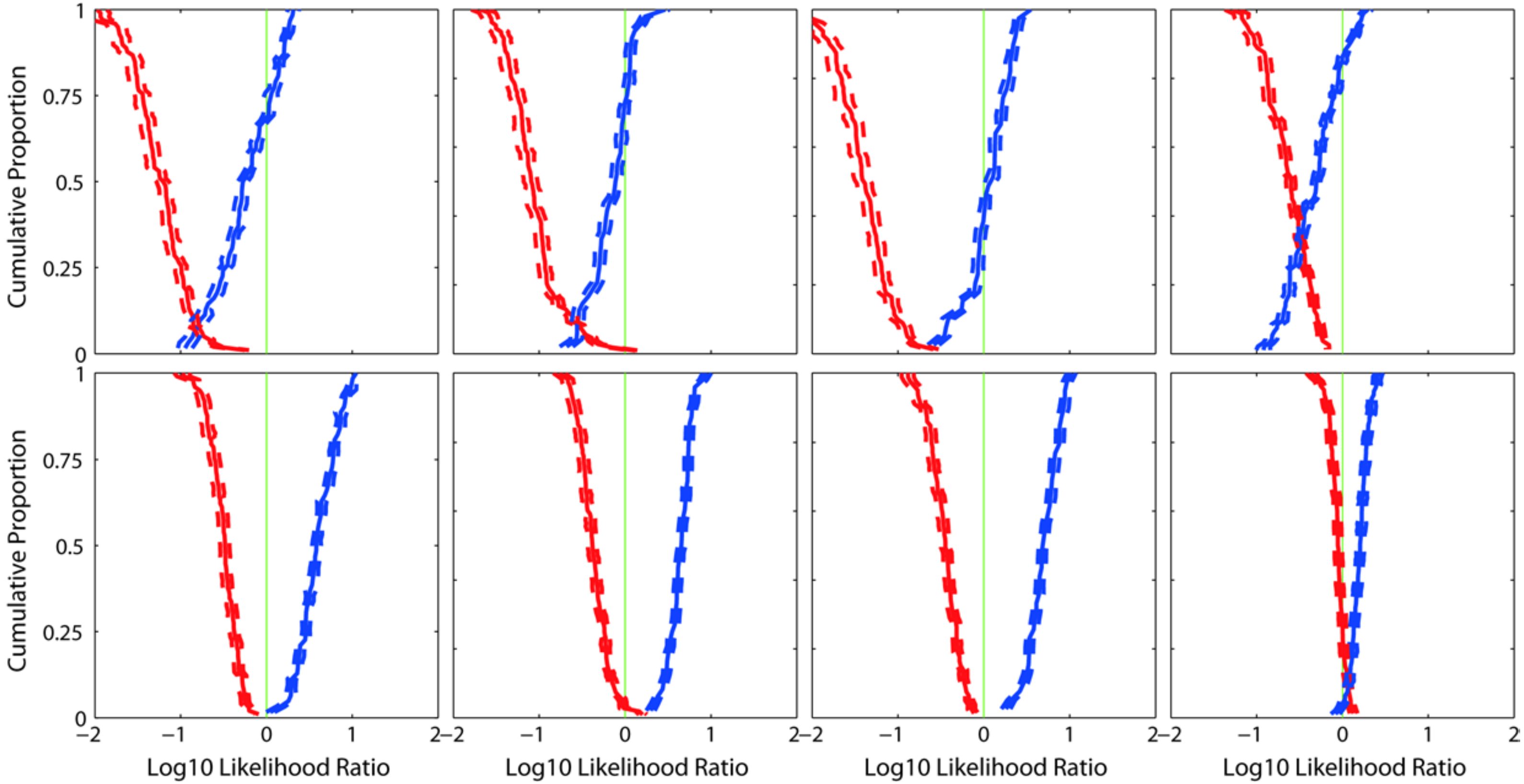
Campbell, W., Sturim, D., Torres-Carrasquillo, P., & Reynolds, D. A. (2008). A comparison of subspace feature-domain methods for language recognition. In: Proc. Interspeech, Brisbane, Australia, pp. 309–312.

Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., Laface, P., 2006. Channel factors compensation in model and feature domain for speaker recognition. In: Proc. Odyssey, San Juan, pp. 1–6.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C., 2007. Compensation of nuisance factors for speaker and language recognition. IEEE Trans. Audio, Speech Lang. Proc. 15 (7), 1969–1978

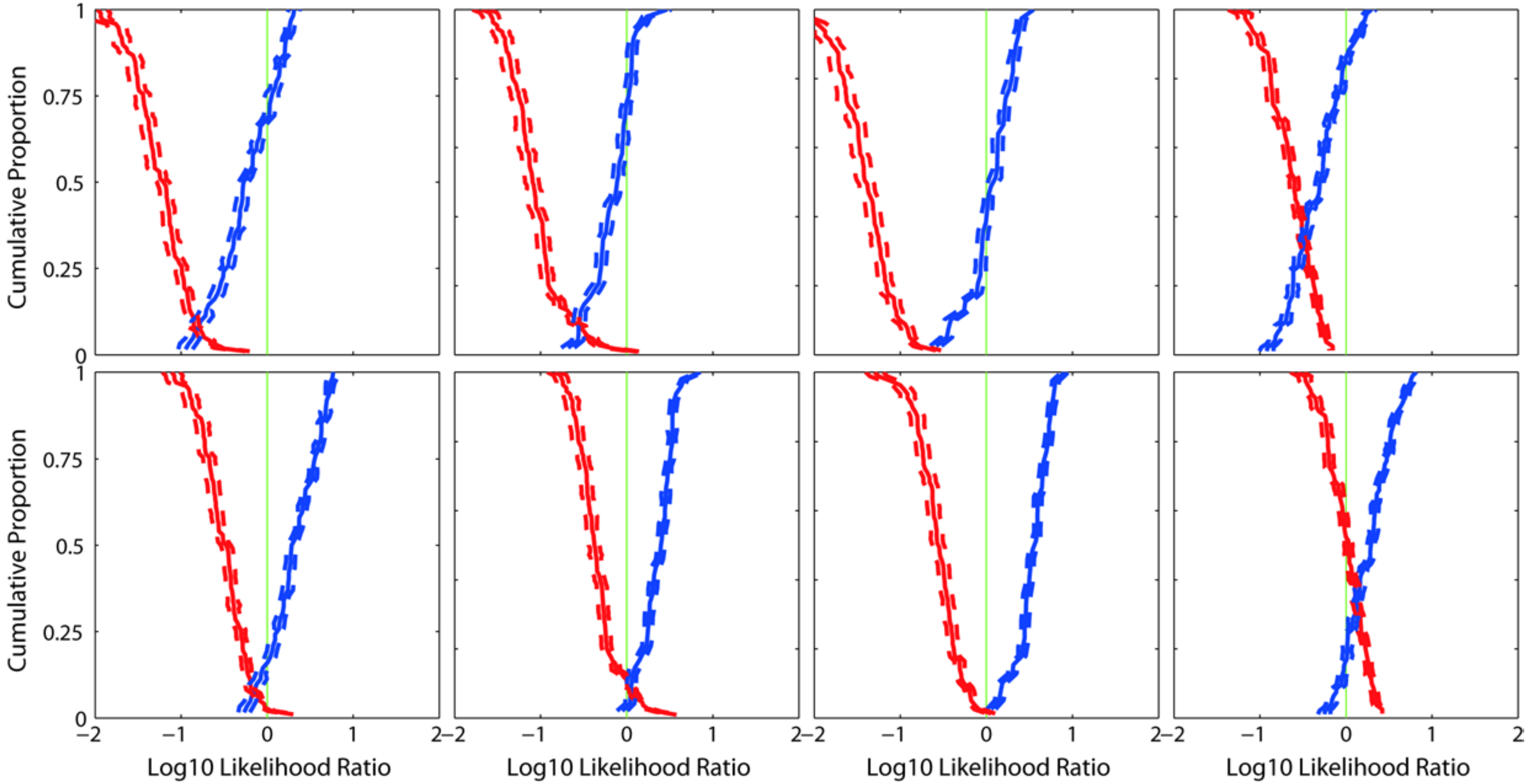
Case study 1: Mismatched distances

- Effect due to mismatch:



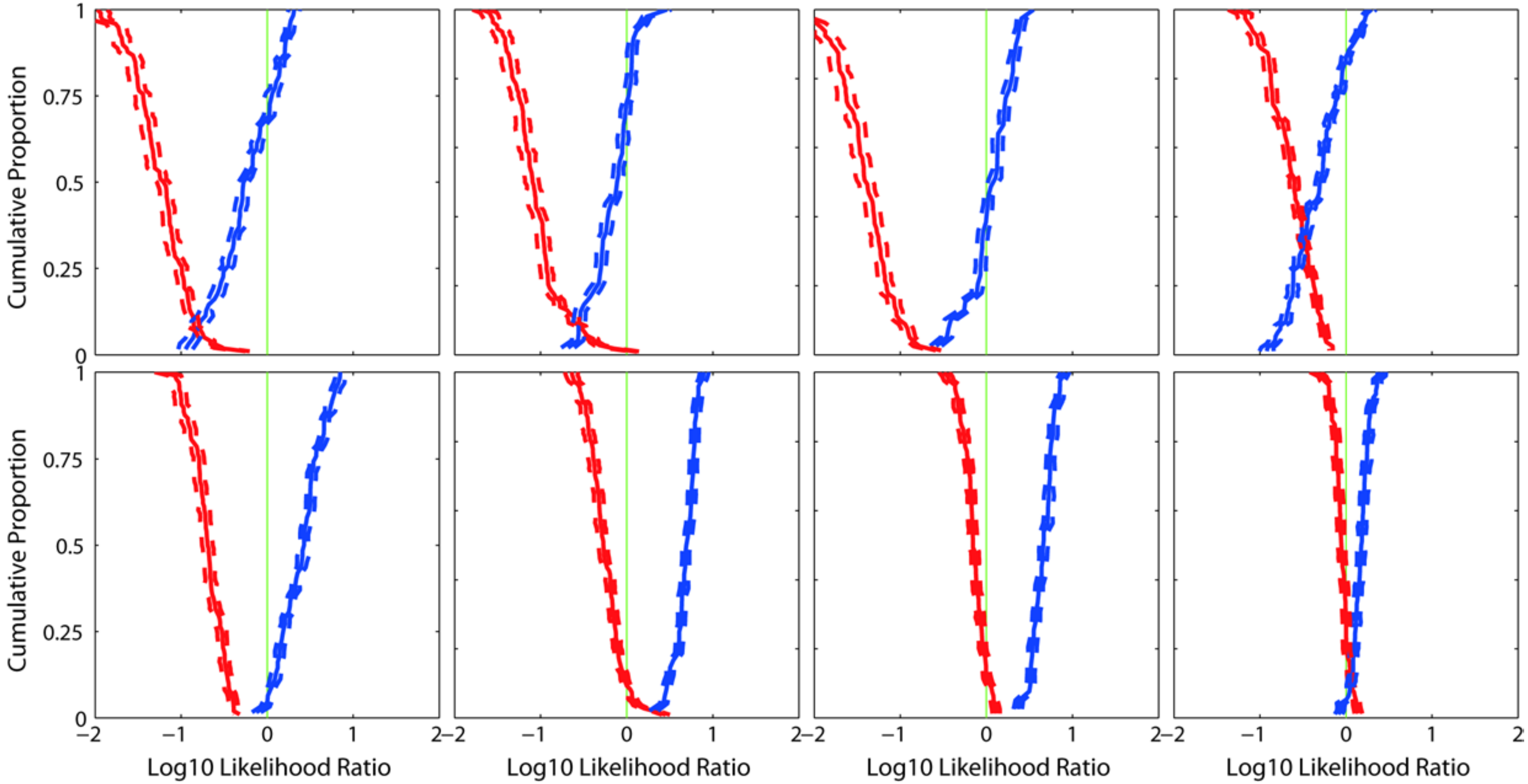
Case study 1: Mismatched distances

- Likelihood-ratio based bias compensation:



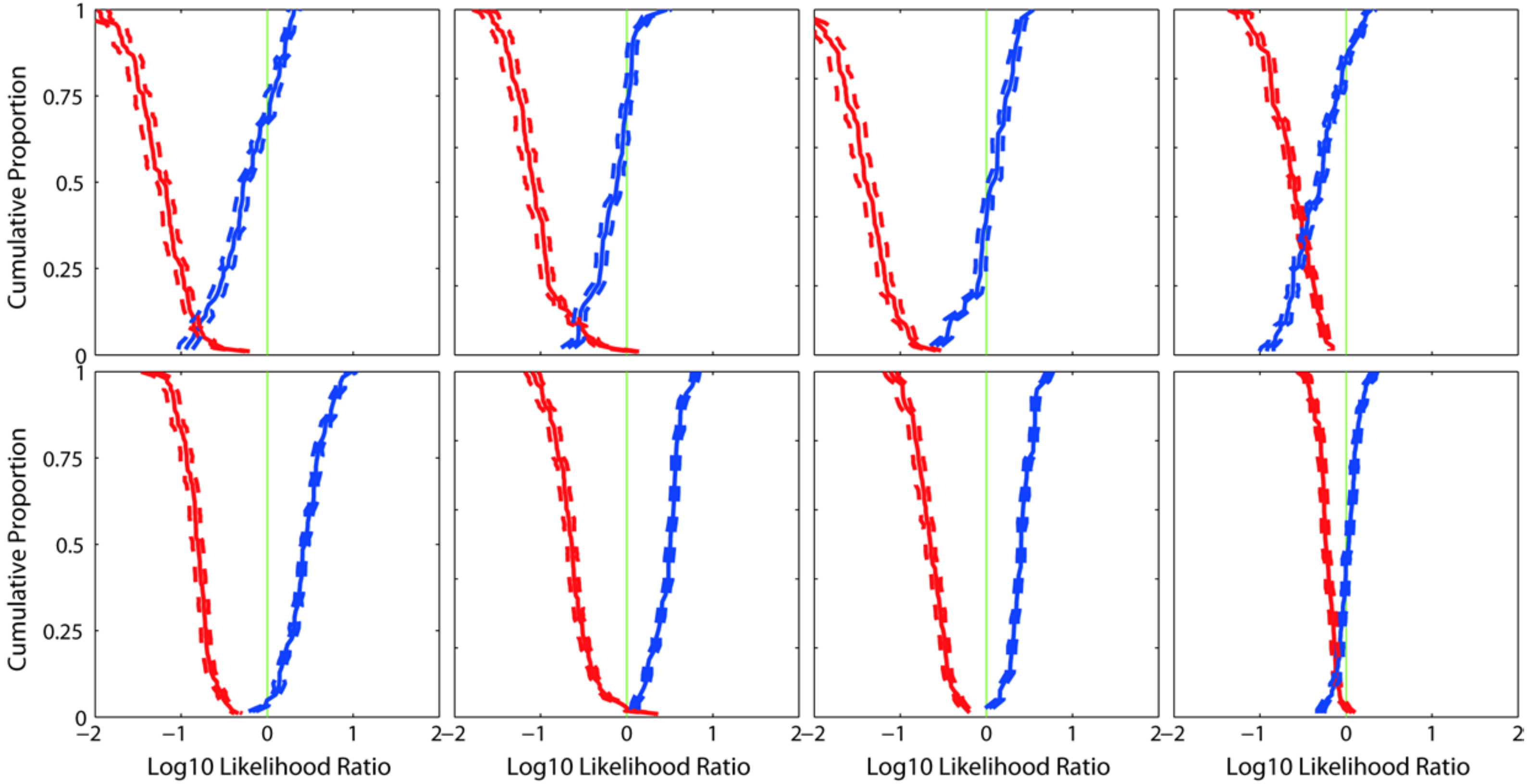
Case study 1: Mismatched distances

- Feature mapping (mean feature vector offset)



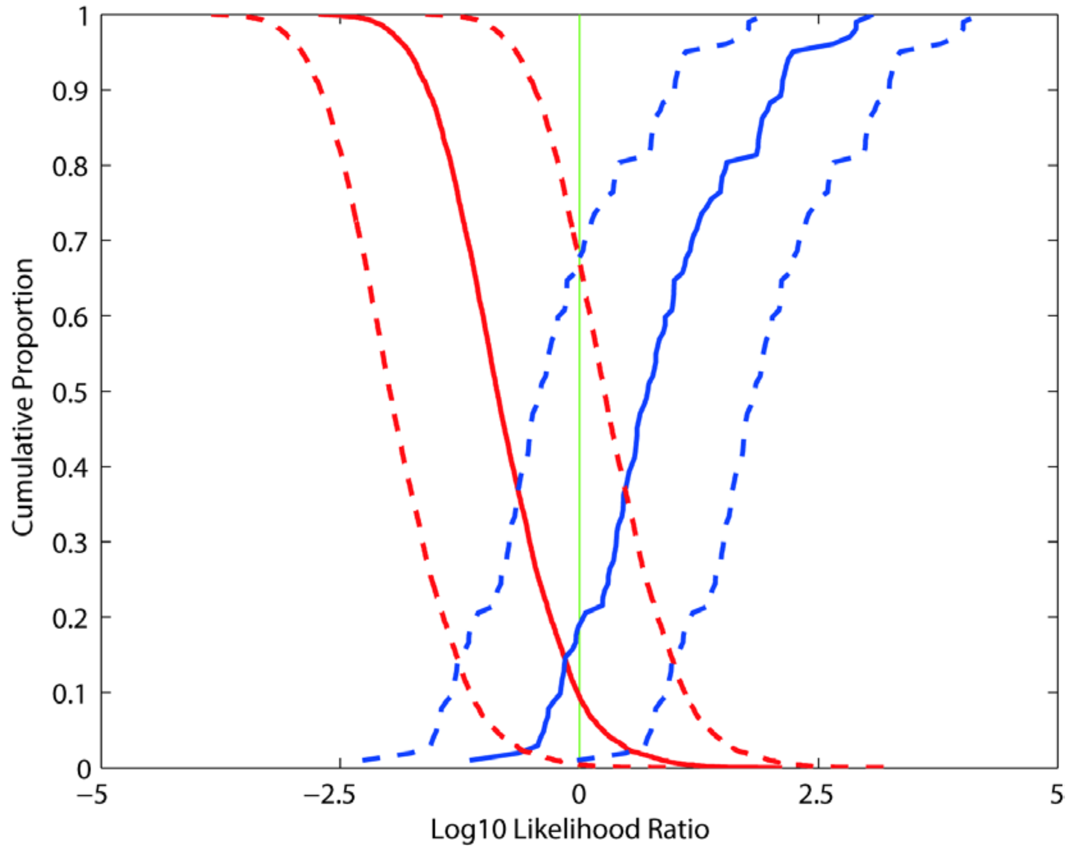
Case study 1: Mismatched distances

- Linear discriminant function analysis transform



Case 2: Australian English males

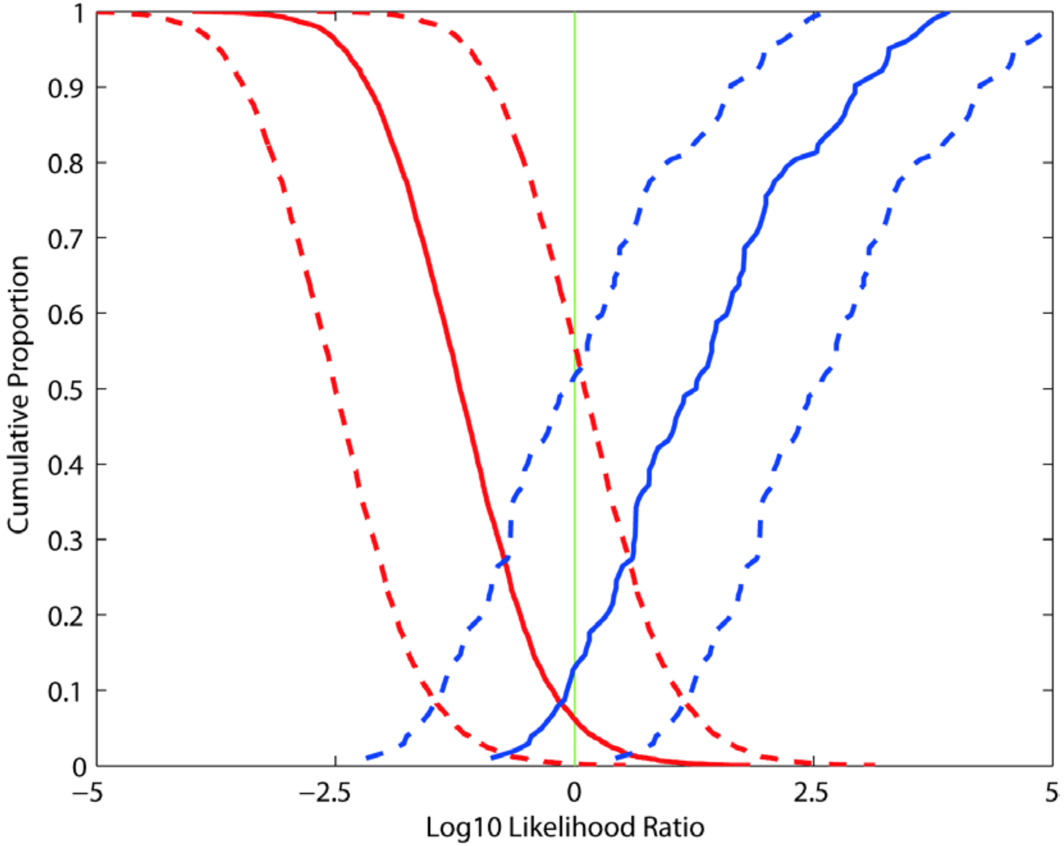
- Baseline (Feature warping)



C_{llr} = 0.442
95%CI = 1.121

Case 2: Australian English males

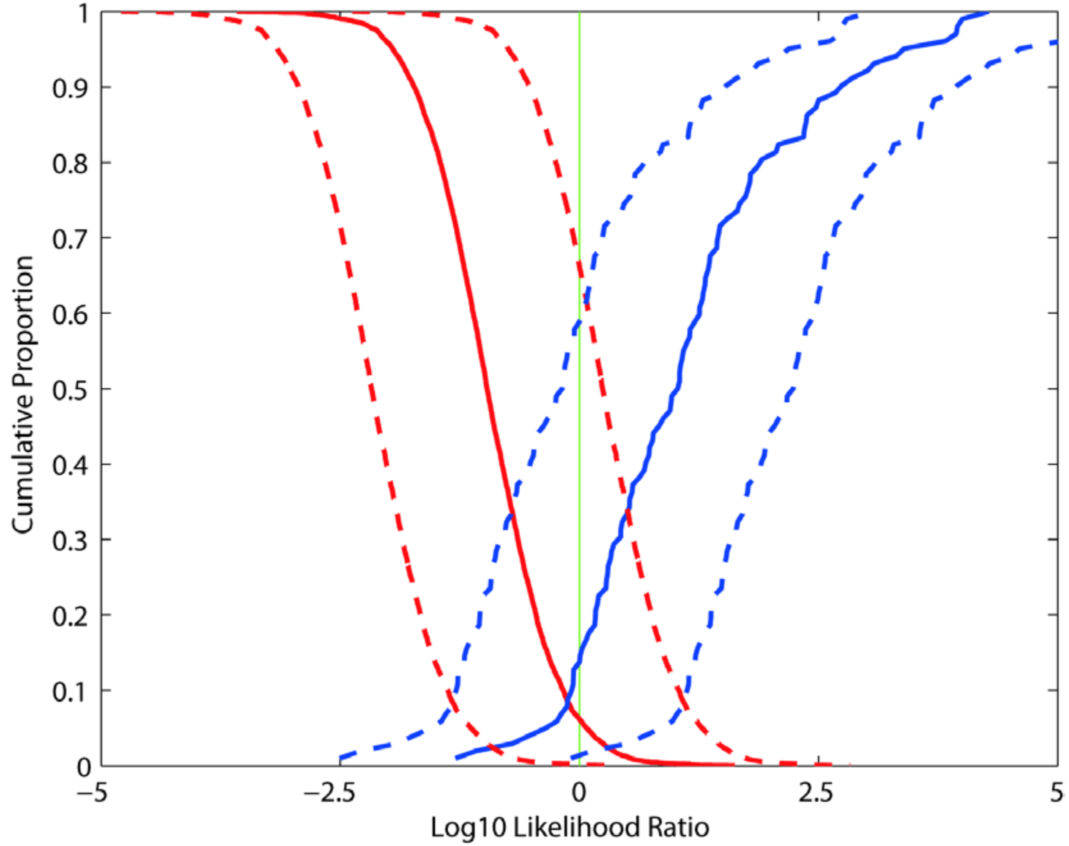
- Probabilistic feature mapping



$C_{llr} = 0.309$
95%CI = 1.304

Case 2: Australian English males

- Feature-domain nuisance attribute projection



C_{llr} = 0.371
95%CI = 1.208

NAP rank = 30