Likelihood ratio calculation in acoustic-phonetic forensic voice comparison: Comparison of three statistical modeling approaches

Ewald Enzinger

University of New South Wales

What is forensic voice comparison (FVC)?

- Task is to assist the court (judge, jury, etc.) to decide whether a recording of a voice of questioned identity was produced by a speaker of known identity or not

- I’m not going to talk about investigative forensic applications
  - e.g. law enforcement agencies searching for a suspect in a database

Paradigm for the evaluation of forensic evidence

- Use of the likelihood ratio framework
  - Logically correct
  - Adopted for DNA in the mid 1990s

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

- Use of relevant data (data representative of the relevant population), quantitative measurements, and statistical models
  - Transparent and replicable
  - Relatively robust to cognitive bias

- Empirical testing of validity and reliability under conditions reflecting those of the case under investigation, using test data drawn from the relevant population

Paradigm for the evaluation of forensic evidence

- Use of the likelihood ratio framework
  - Logically correct
  - Adopted for DNA in the mid 1990s

$$\boxed{\mathrm{LR} = \frac{p(E|H_p)}{p(E|H_d)}}$$

- Use of relevant data (data representative of the relevant population), quantitative measurements, and statistical models
  - Transparent and replicable
  - Relatively robust to cognitive bias

- Empirical testing of validity and reliability under conditions reflecting those of the case under investigation, using test data drawn from the relevant population

Acoustic-phonetic-statistical FVC

- Manual segmentation
- Quantitative measurement of acoustic-phonetic properties
  - Formants / formant trajectories
  - Fundamental frequency
  - Cepstral coefficients
  - …
- Statistical modeling of quantitative measurements
  - Assess “similarity” and “typicality” in LR calculation

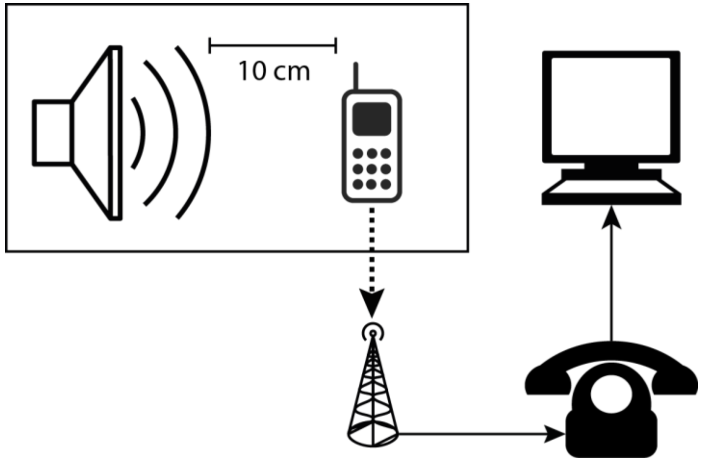
Statistical modeling

- Multivariate kernel density (MVKD)
  - “standard” model used in acoustic-phonetic FVC research
  - Problems with higher-dimensional data, data sparsity
- Principal component analysis kernel density (PCAKLR)
  1. Obtains decorrelating transform using PCA
  2. Computes LR as the product of univariate kernel-density based likelihood ratios of the projected features
- Multivariate normal model (MVN)
  - More parsimonious model

Data

- 60 female Standard Chinese speakers

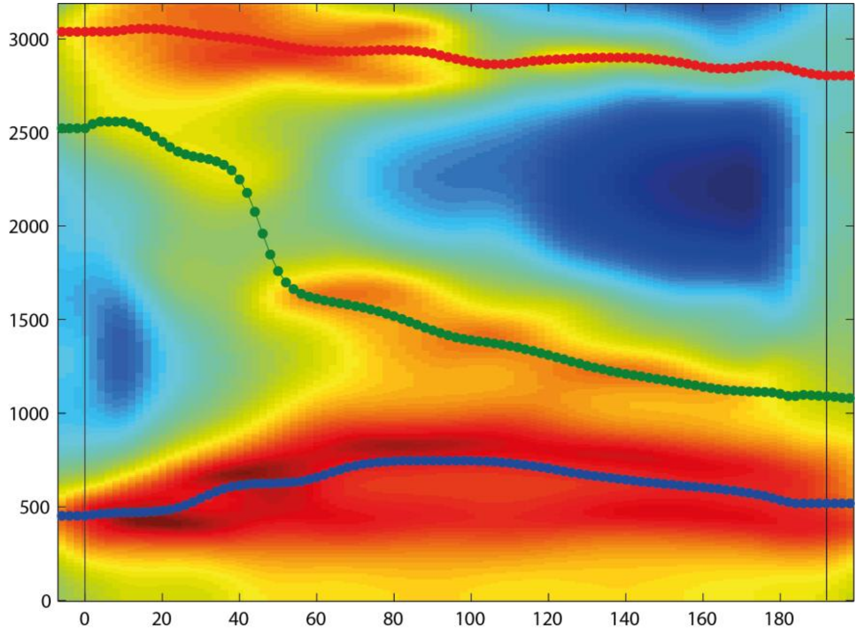
  Available: `http://databases.forensic-voice-comparison.net/`
- Two recording sessions separated by 2-3 weeks
- Information-exchange task over the telephone
- Channels:
  - High-quality
  - Mobile-to-landline transmission
- Split into 3 groups of 20 speakers:
  - background set
  - development set
  - test set



Quantitative measurement

- Manually marked /iau/ tokens in stressed positions
- Human-supervised formant-trajectory measurement
  (FORMANTMEASURER, Morrison & Nearey)
- 0th through 4th discrete cosine transform (DCT)
- Coefficients of F2 and F3

➢ 10-dimensional features



Baseline automatic MFCC GMM-UBM system

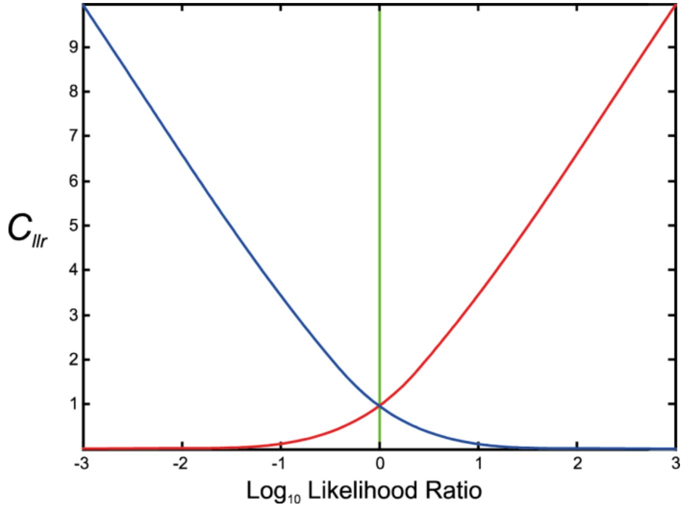
- Entire speech-active portion of recordings
- 16 Mel frequency cepstral coefficients (MFCCs) + Δ
- Feature warping
- Gaussian mixture model – universal background model
- Logistic-regression calibration/fusion

- Evaluation with respect to improvement/degradation in performance of fused system relative to baseline system

Evaluation measures

- Validity / Accuracy:
  - Log-likelihood ratio cost (C_{llr}) metric

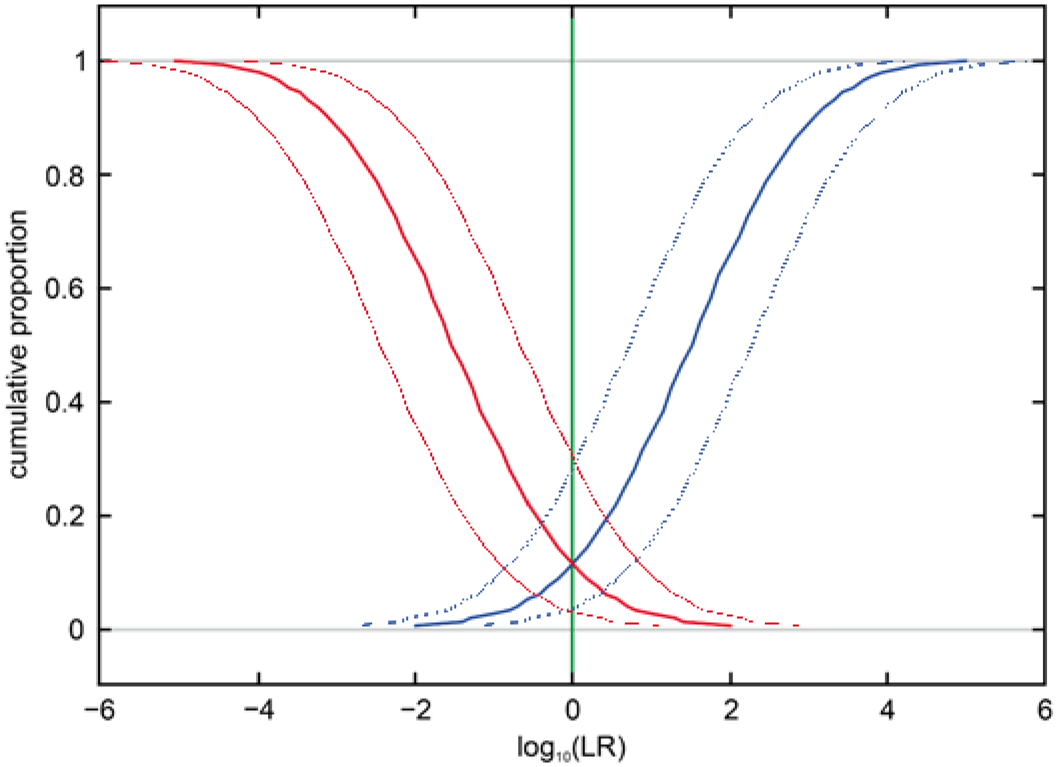
$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{H_p}}\sum_{i=1}^{N_{H_p}}\log_2\left(1+\frac{1}{LR_{i,H_p}}\right)+\frac{1}{N_{H_d}}\sum_{j=1}^{N_{H_d}}\log_2\left(1+LR_{j,H_d}\right)\right)$$



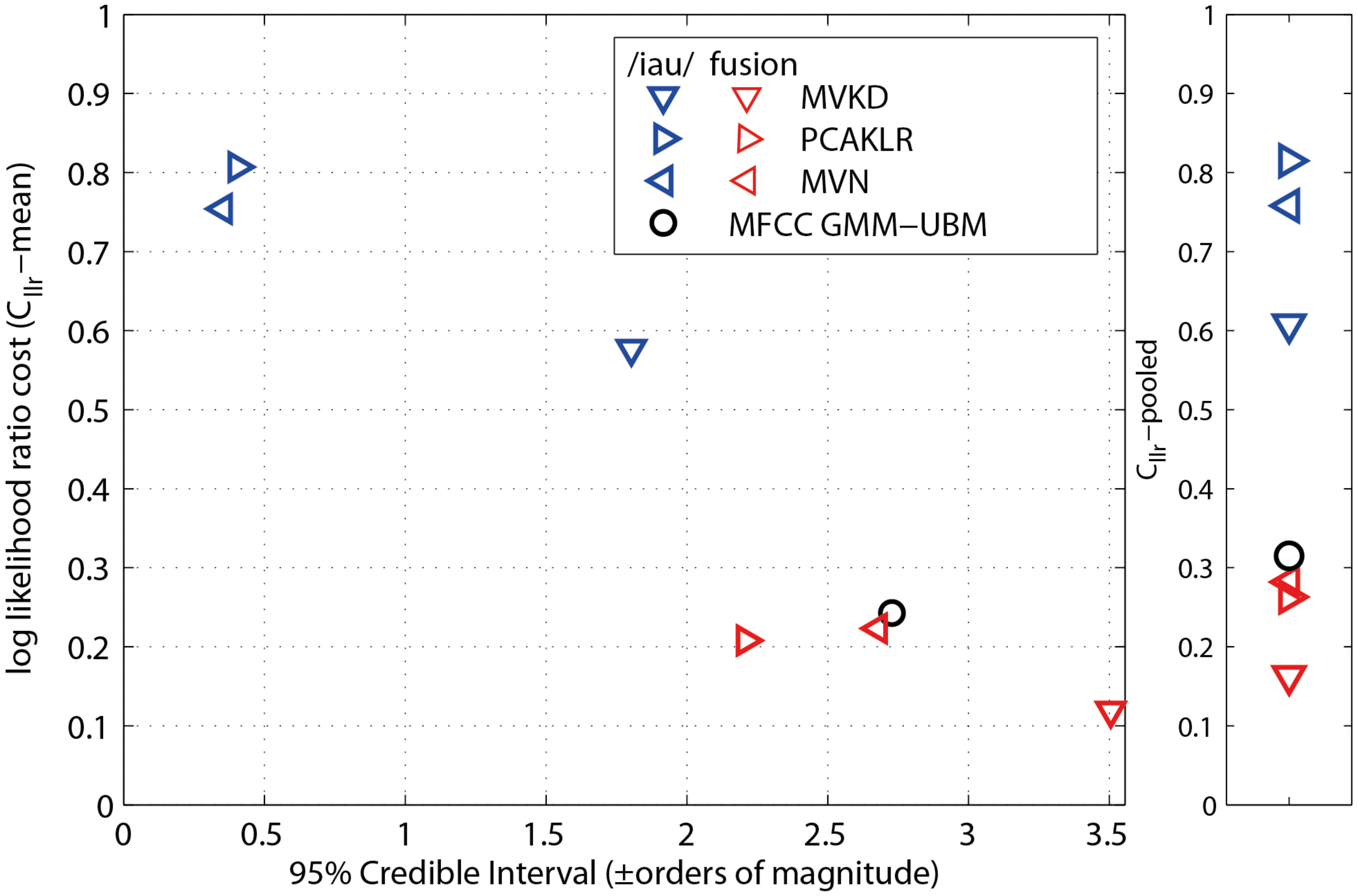
- Reliability / Precision
  - Multiple comparisons per speaker pair (using different recordings)
  - Estimate 95% credible interval

Evaluation measures

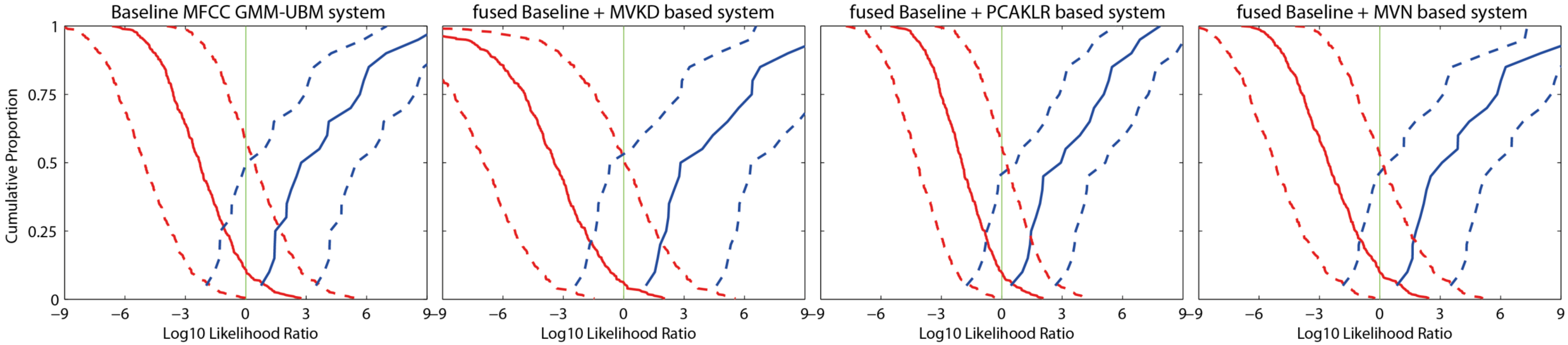
- Graphical presentation using Tippett plots
  - Different-speaker LRs
  - Same-speaker LRs



Results – Validity and reliability



Results – Tippett plots



MFCC GMM-UBM (Baseline)

Fusion Baseline + MVKD system

Fusion Baseline + PCAKLR system

Fusion Baseline + MVN system

Monte Carlo simulation

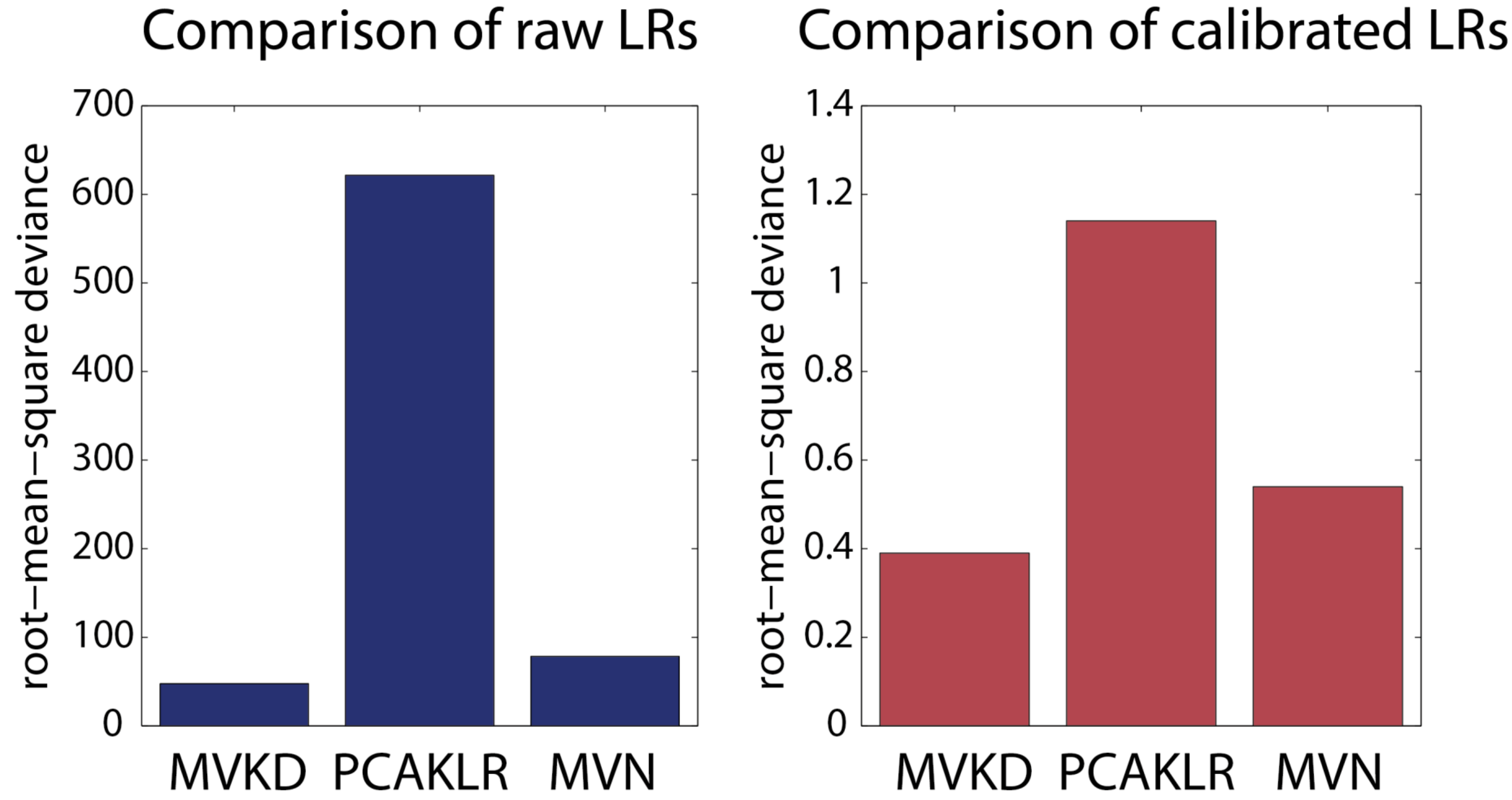
- In practice, the true distribution for a given population is not known

➢ Comparison of LR estimate with “true” LRs in Monte-Carlo simulation

1. Generate sets of measurements for 1000 simulated speakers
2. Calculate “true” LRs based on specified distributions
3. Calculate LRs using MVKD, PCAKLR, MVN
4. (Optional:) Calibrate LRs

- Evaluation measure:
    - Root-mean-square deviation between estimated and “true” LRs

Results – Monte Carlo simulation



Comparison of raw LRs

Comparison of calibrated LRs

Conclusions

- Multivariate kernel density (MVKD):
  - Best overall performance on real data
  - Lowest RMS deviation from “true” LRs in Monte-Carlo simulations
  - ➢ Provides empirically best performance

- Caveats:
  - Only single phonetic unit (/iau/)
  - Only single type of features (formant trajectory DCTs)
  - Only female speakers, one speaking style, specific mismatch condition

Thanks

http://entn.at/

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

http://forensic-evaluation.net/

Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*)

Organizers: Geoffrey Stewart Morrison & Ewald Enzinger

- Evaluation of forensic voice comparison systems
- Training and test data reflect the conditions of real case
- Operational and research laboratories are invited to participate
- Results will be published in a Virtual Special Issue of *Speech Communication*

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