Testing the validity and reliability of forensic voice comparison based on reassigned time-frequency representations of Chinese /iau/

Ewald Enzinger

School of Electrical Engineering & Telecommunications, The University of New South Wales, Sydney, Australia

Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

Paradigm for evaluation of FVC evidence

- Likelihood-ratio framework:
  - Statement of strength of the evidence as an answer to a specific question

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

- Quantitative measurements, statistical models, databases representative of the relevant population

- Testing of validity and reliability under conditions reflecting those of the case

Introduction

- Fulop & Disner (2007, 2009):
  - Pruned T-F-reassigned spectrograms of short vowel segments ([æ], [a] etc.)
  - visual comparison of spectrograms by human experts (“voiceprint”)
  - Fulop (2011): U.S. Patent 8,036,891 B2
- Fulop & Kim (2013):
  - Quantitative approach
  - Automatic SVM-based closed-set identification
  - 24 enrolled speakers, 6 test segments

TF reassigned spectrograms

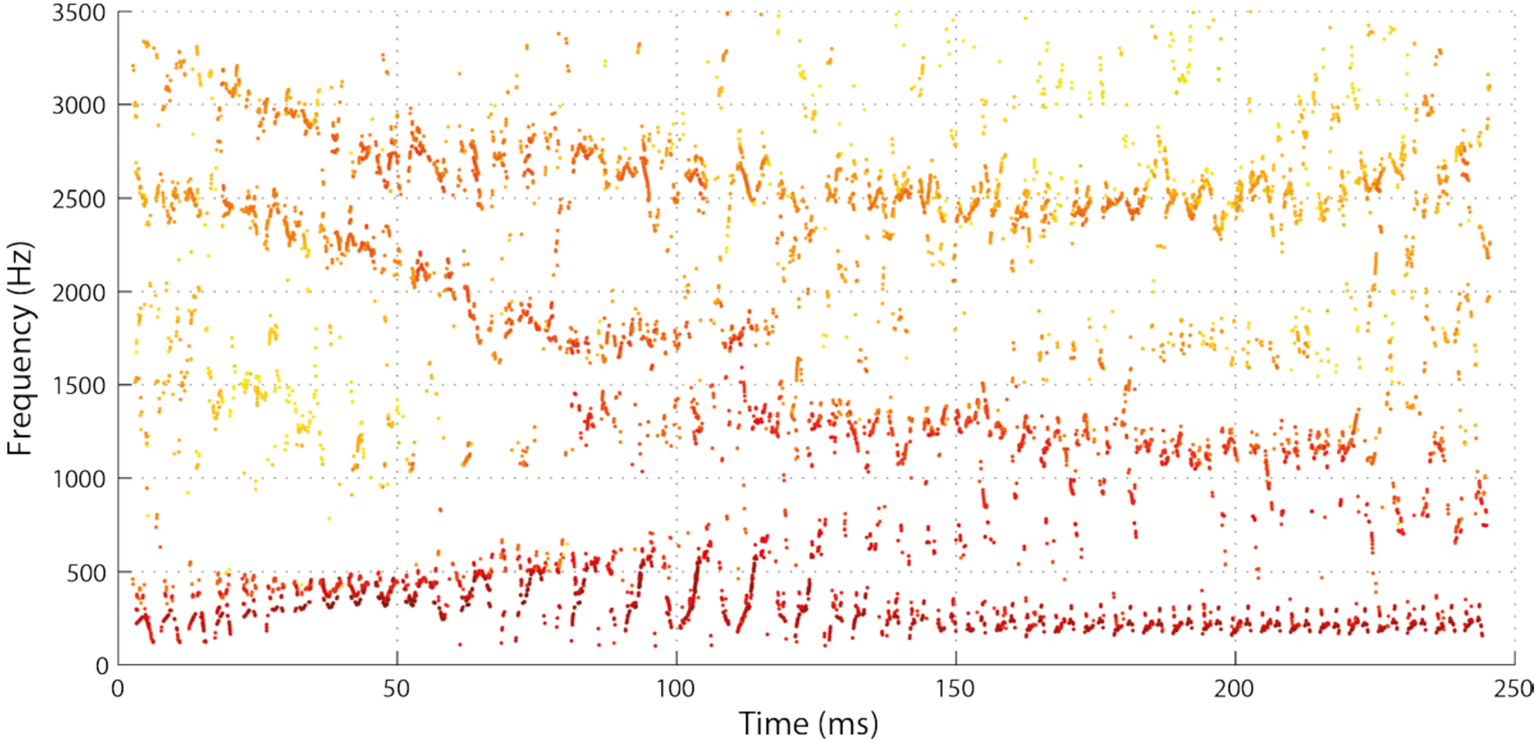
- Short-time Fourier transform of /iau/
- Channelized Instantaneous Frequency (CIF)
- Local Group Delay (LGD)

$$
\begin{aligned}
\mathrm{CIF}(\omega, T) &= \frac{\delta}{\delta T} \arg(X_h(\omega, T)) \\
\mathrm{LGD}(\omega, T) &= \frac{\delta}{\delta \omega} \arg(X_h(\omega, T))
\end{aligned}
$$

- “Reassign” T-F magnitudes to locations corresponding to local center of gravity

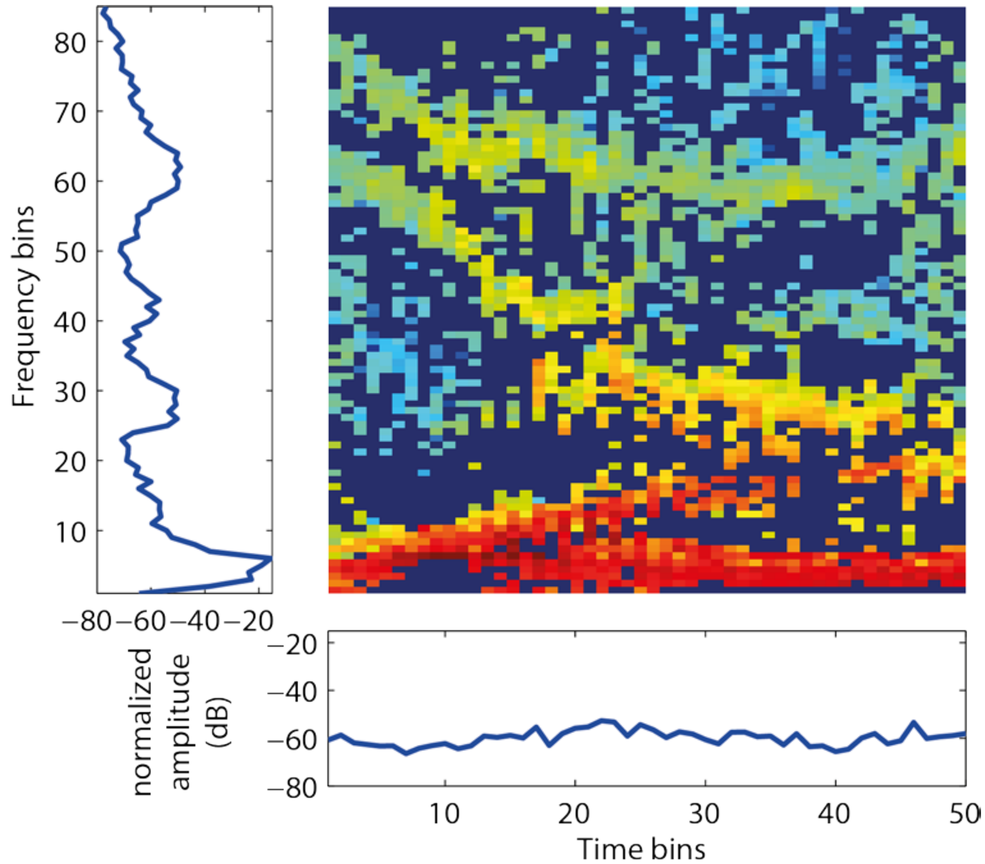
TF reassigned spectrograms

- Pruning (threshold) to reduce noise/artefacts
  - Based on second-order mixed partial derivative (Nelson, 2001)
  - Set to retain line components and impulses



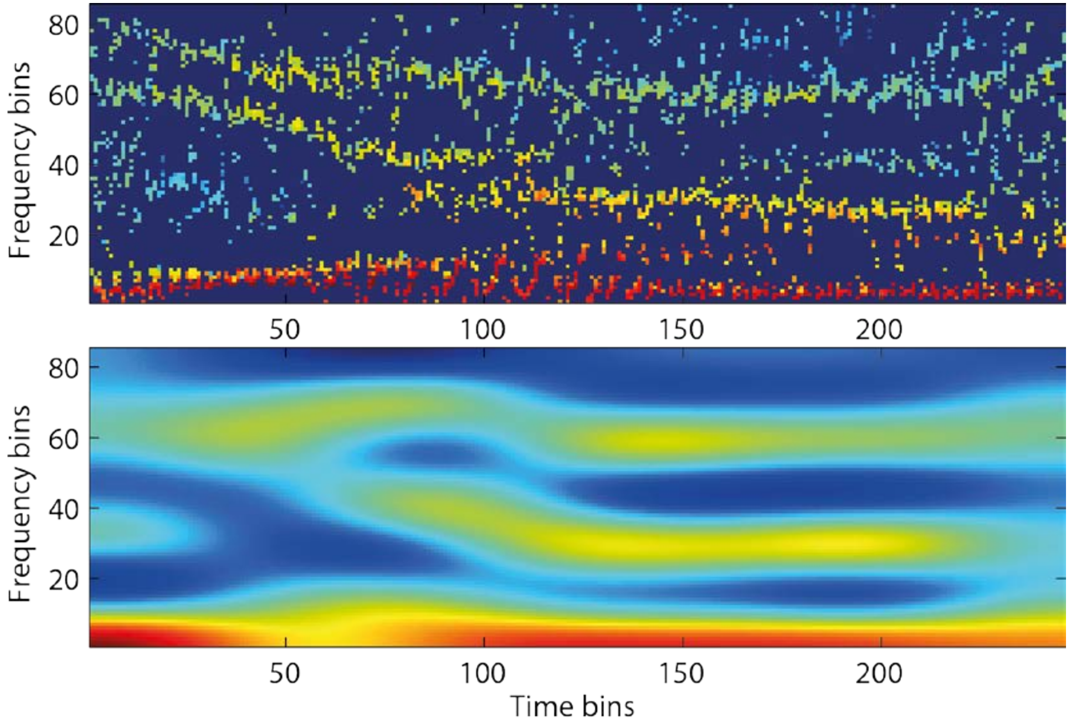
TF feature representation – **TFR AVG**

- Fulop & Kim (2013): Feature representation based on discretization using a coarse grid
  - 50 time bins
  - 85 frequency bins
- Dimensionality reduction via PCA
  - 10 time features
  - 20 frequency features



TF feature representation – **TFR DCT**

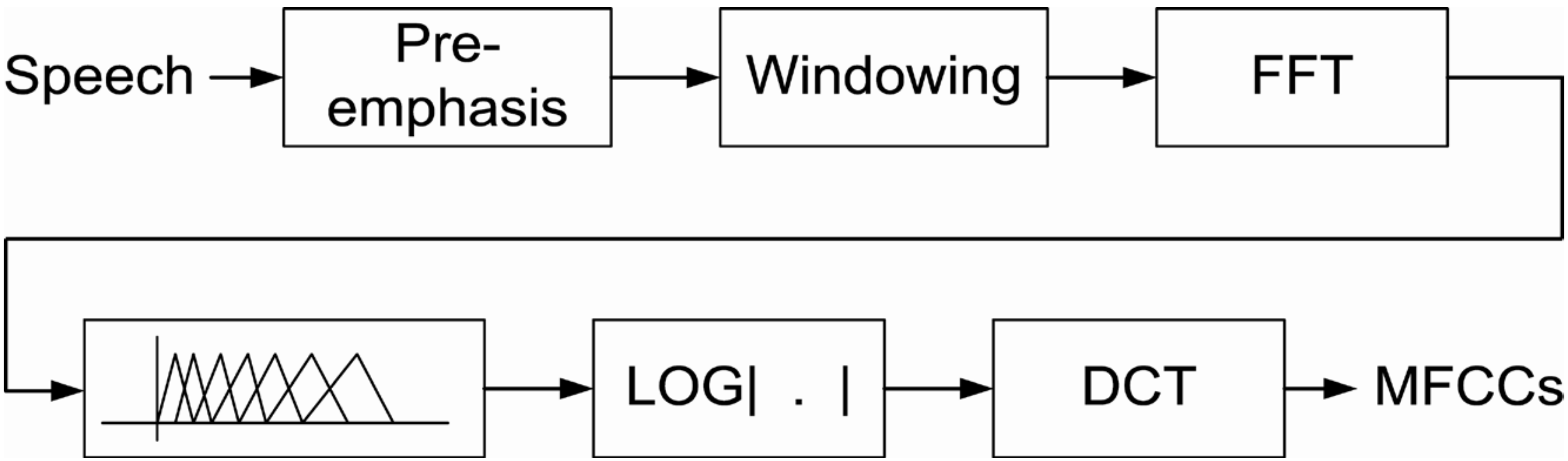
- Chinese /iau/ triphthong:
  - Significant correlation over time and frequency
  - 2D Discrete cosine transform (DCT)



lower-order 7 x 7 coefficients

Feature representation – **MFCC-on-/iau/**

- Mel frequency cepstral coefficients (MFCC)
  - Common feature in FVC / speaker recognition
  - Extracted from /iau/ triphthong tokens
  - 16 MFCC + 16 Delta (Δ) coefficients



Likelihood ratio calculation

- Score obtained using Gaussian mixture model-Universal background model (GMM-UBM) approach

$$\lambda = (p_i, \mu_i, \Sigma_i)_{i=1,\ldots,M} \qquad \mathrm{s} = \frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{p(x_j \mid \lambda_{\mathrm{suspect}})}{p(x_j \mid \lambda_{\mathrm{UBM}})}\right)$$

- Logistic regression calibration and fusion
- Baseline automatic FVC system
  - Entire speech-active portion of recording
  - 16 MFCC + 16 delta (Δ) coefficients
  - 1024 Gaussian mixture components (UBM)

Data

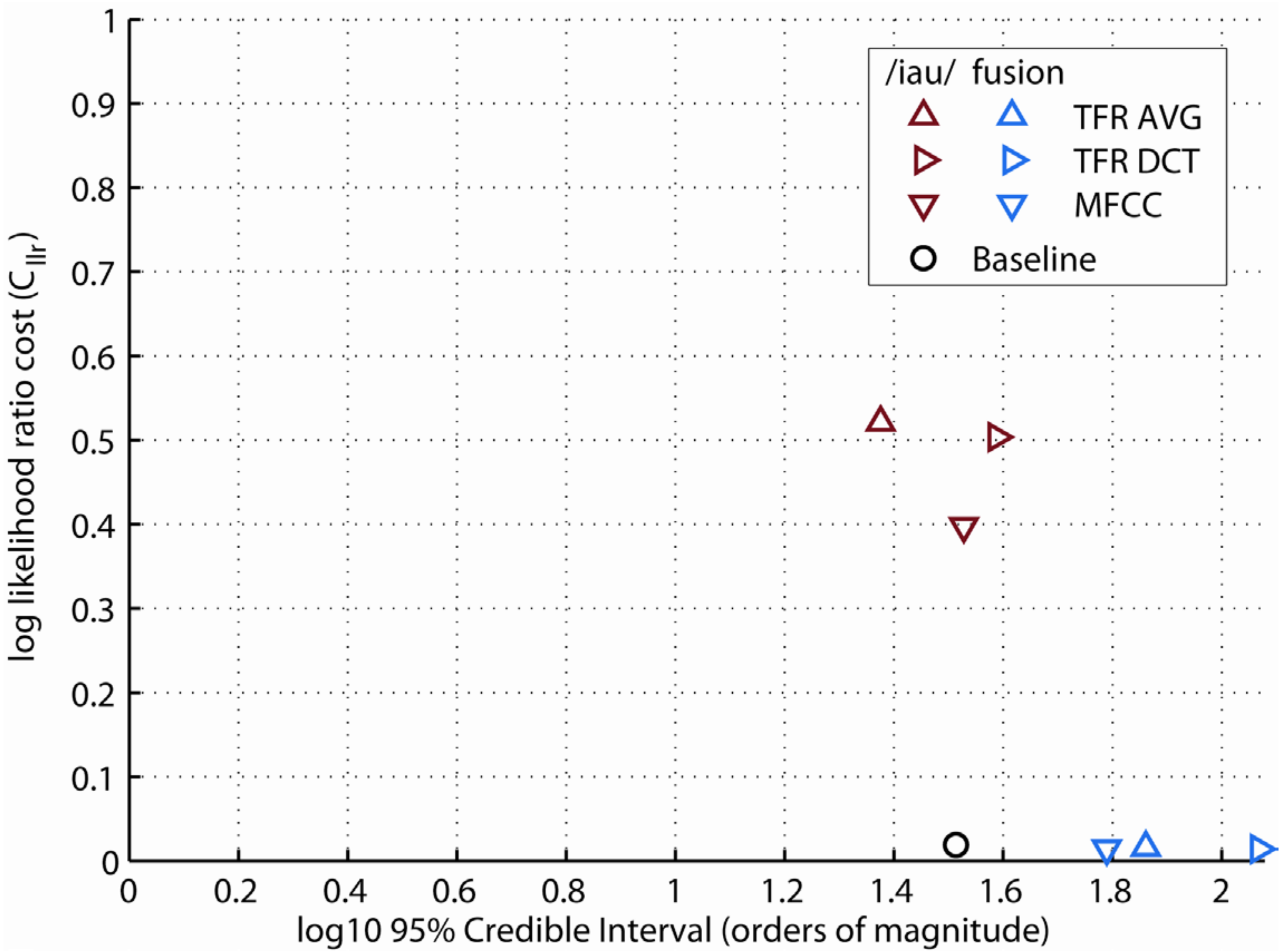
- 60 female Standard Chinese speakers
- Split into 3 groups of 20 speakers
  - background set
  - development set
  - test set
- Manually marked /iau/ triphthongs
- Information-exchange task over telephone
- High quality and mobile-to-landline data
- Two recording sessions separated by 2–3 weeks

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

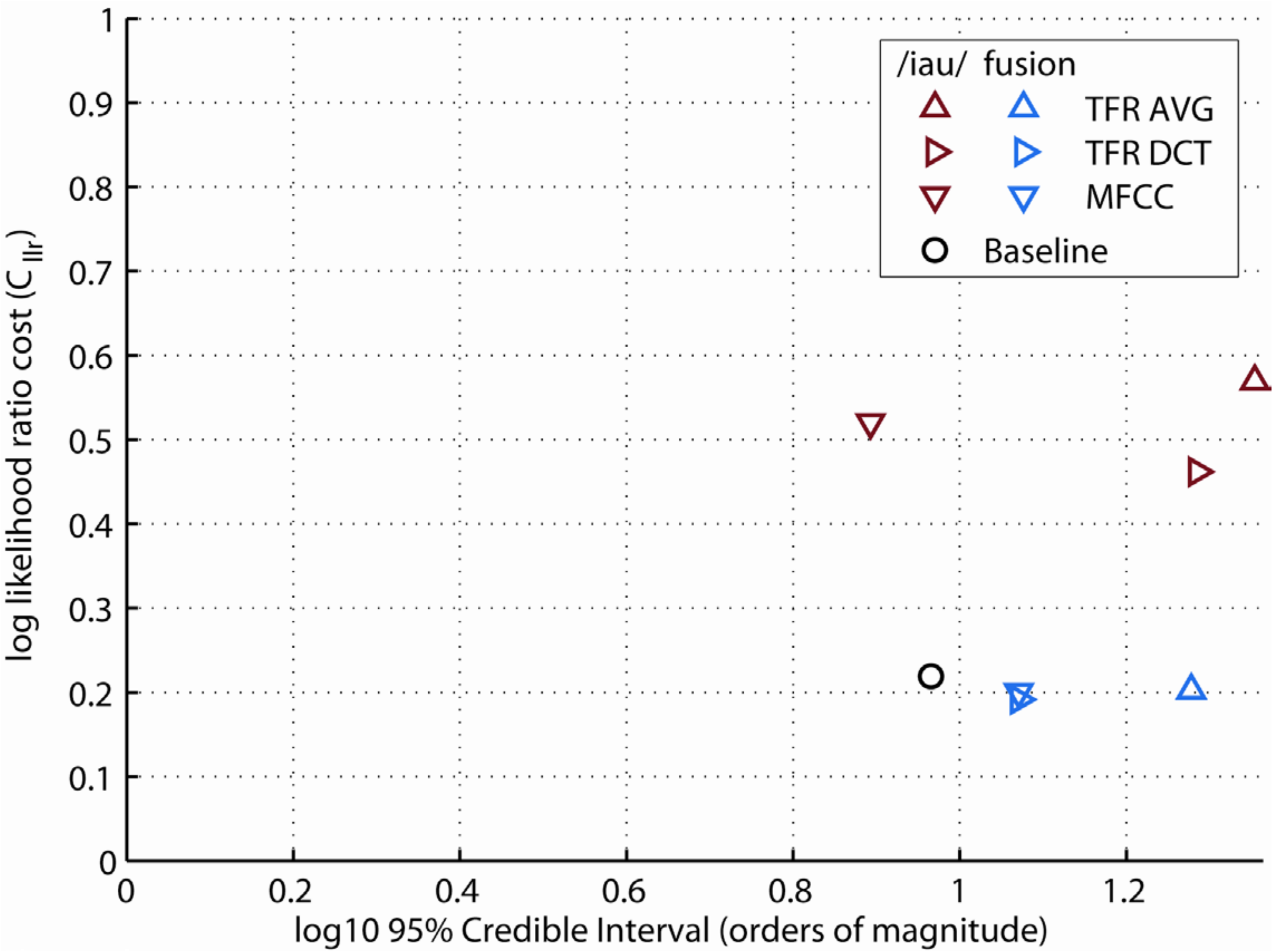
Evaluation

- Validity / Accuracy
  - log-likelihood ratio cost (C_{llr}) metric
- Reliability / Precision
  - 95% credible interval (Morrison, 2011)
- Conditions:
  - High-quality v high-quality
  - Mobile-to-landline v mobile-to-landline
  - High-quality v mobile-to-landline

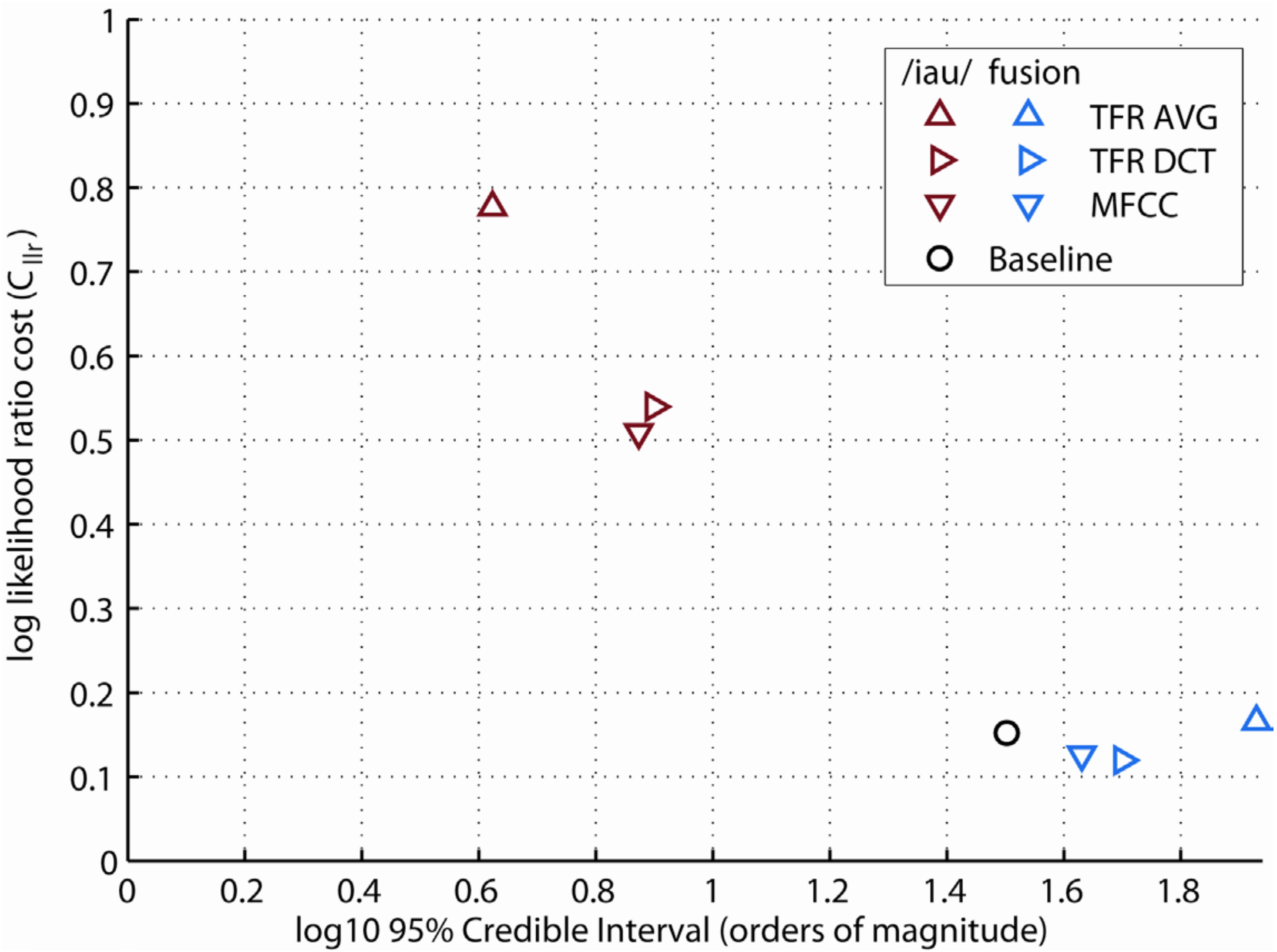
Results – high-quality v high-quality



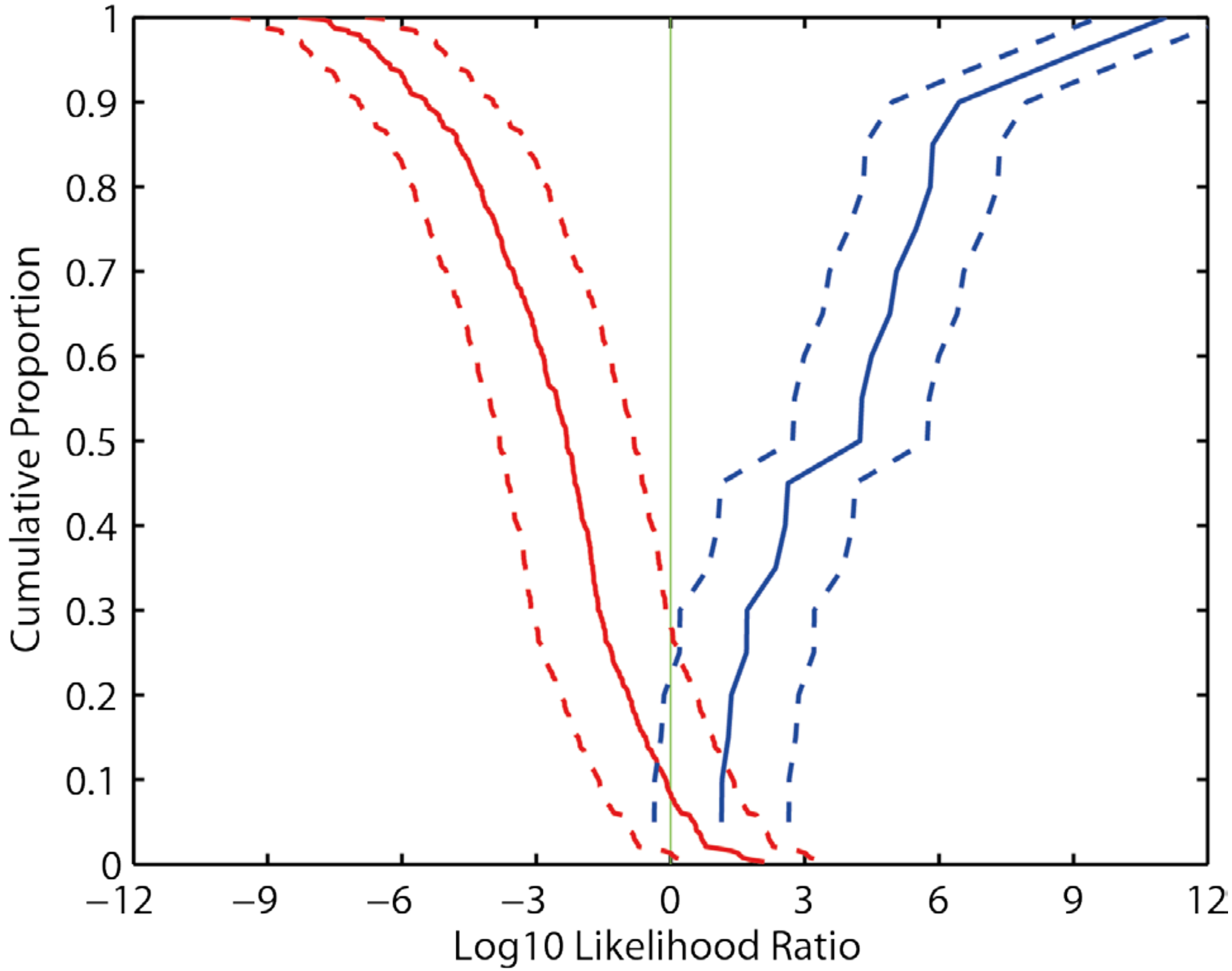
Results – mobile-to-landline v mobile-to-land



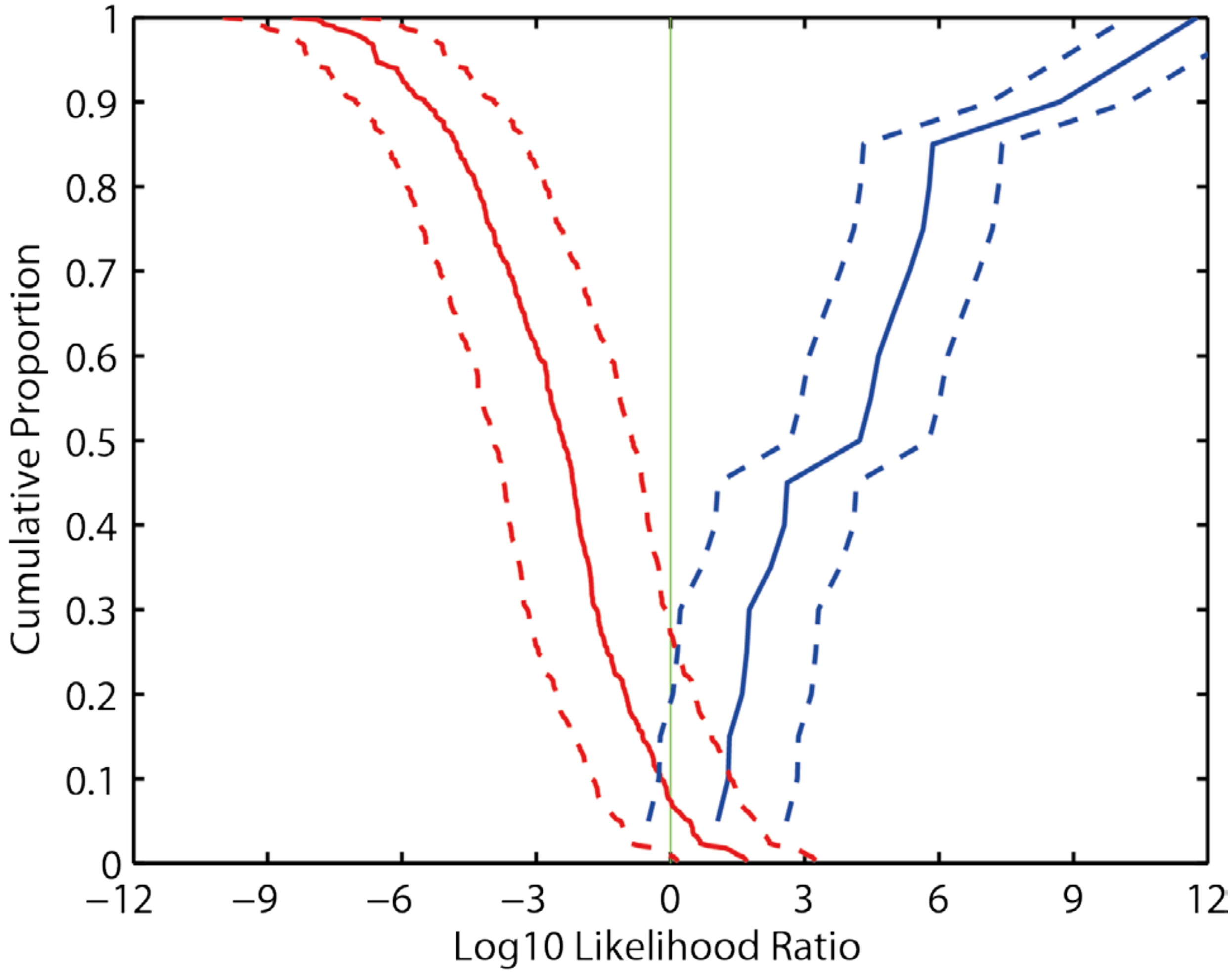
Results – mobile-to-landline v high-quality



Tippett plot – Baseline system



Tippett plot – Fusion Baseline + TFR DCT



Conclusion

- High-quality v high-quality
  - no substantial improvement
- Mobile v mobile, mobile v high-quality
  - Improvement in validity, reliability deteriorates
  - TFR DCT improves upon TFR AVG
  - MFCC-on-/iau/ similar or slightly better
- Caveat:
  - Results give only an indication of performance (not tested: background noise, reverberation, ..)
  - Testing on per-case basis

Thank You!!

References

Fulop, S. A. & Disner, S. F. (2007). The reassigned spectrogram as a tool for voice identification. In: Proc. ICPhS XVI, Saarbrücken, Germany, pp. 1853–1856.

Fulop, S. A. & Disner, S. F. (2009). Advanced time-frequency displays applied to forensic speaker identification. In: Proceedings of Meetings on Acoustics, vol. 6, 2009, paper 060008.

Fulop, S. A. & Kim, Y. (2013). Speaker identification made easy with pruned reassigned spectrograms. In: Proceedings of Meetings on Acoustics, vol. 19, paper 055068.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. Science & Justice, 51, 91–98. doi:10.1016/j.scijus.2011.03.002

Nelson, D. J. (2001). Cross-spectral methods for processing speech. J. Acoust. Soc. Am., 110, 2575–2592.