A first attempt at compensating for effects due to recording-condition mismatch in formant-trajectory-based forensic voice comparison

Ewald Enzinger

SCHOOL OF ELECTRICAL ENGINEERING & TELECOMMUNICATIONS

UNSW
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

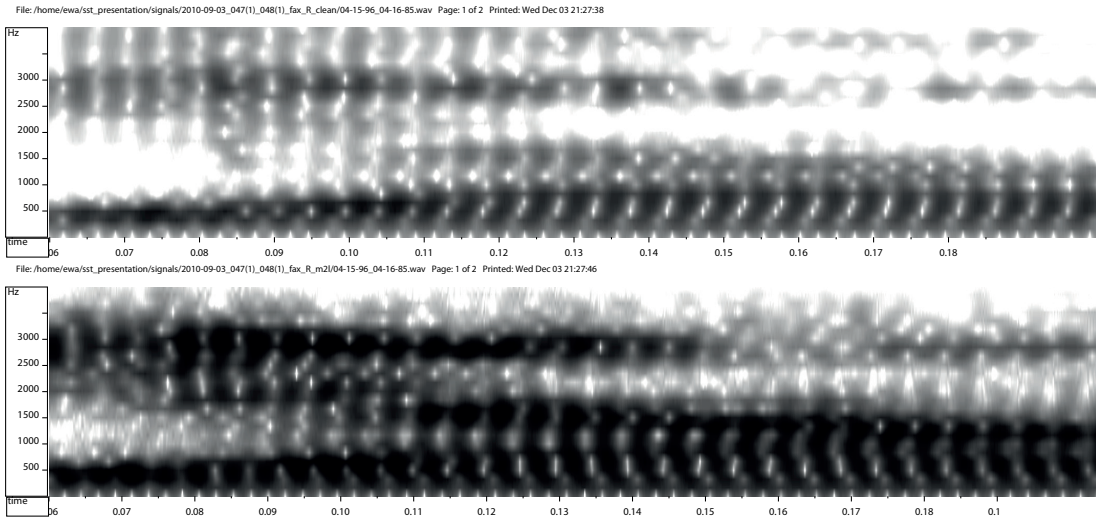
SST 2014
Christchurch, New Zealand

Recording-condition mismatch

Common scenario in FVC:

- Offender recording obtained from telephone call
- Direct-microphone recording of voice of suspect during police interview

Transmission/recording systems affect speech signals



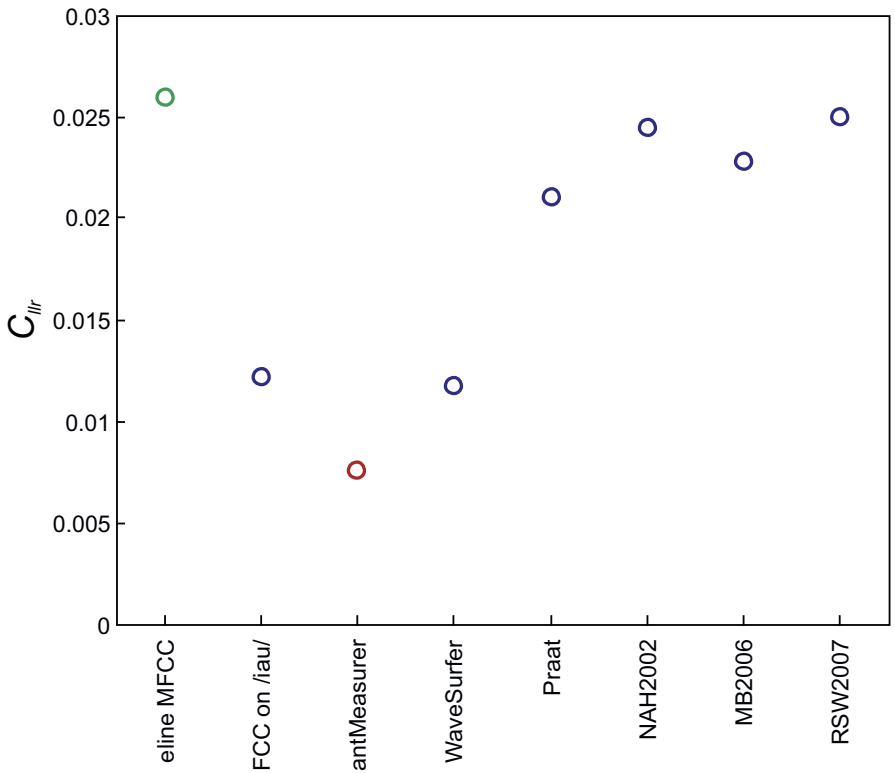
- Expected increase in variability in measurements
- Decrease in ratio of between and within-speaker variability

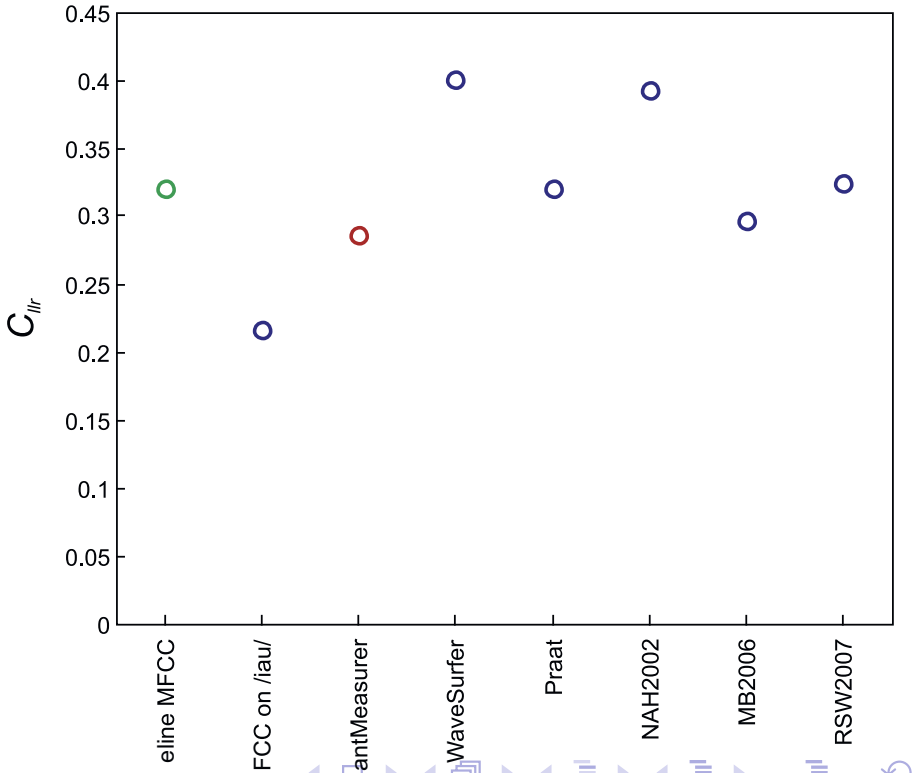
  ➡ Decrease in FVC performance

Telephone transmission and formant-based FVC

Prior studies:

- Average differences in formant frequencies of up to 23% in landline (Künzel, 2001) and 29% in mobile-telephone-transmitted speech (Byrne & Foulkes, 2004)
- Effect of mismatch on performance of formant-trajectory based FVC (Zhang, Morrison, Enzinger, & Ochoa, 2013):





Compensating for recording-condition mismatch

- First attempt at compensation for mismatch between suspect and offender recording conditions in formant-trajectory-based FVC
- Based on statistical distribution of measurements made under respective conditions

Caveat: Most FVC recordings have mismatched recording conditions due to many different factors

- Transmission/recording effects
  - Microphone characteristic, landline telephone bandpass, Mobile telephone codecs, Lossy compression algorithms
- Background noise
  - Heating, air conditioning, ventilation, vehicle noise, etc.
- Reverberation

This study: Focus on mobile-to-landline v high-quality mismatch

Data

- 60 female Standard Chinese speakers
- Transmission/recording conditions:
  - high quality audio
  - mobile-to-landline
- Two recording sessions separated by 2–3 weeks
- Information-exchange task over the telephone
- Split into 3 groups of 20 speakers
  - background database
  - development set
  - test set

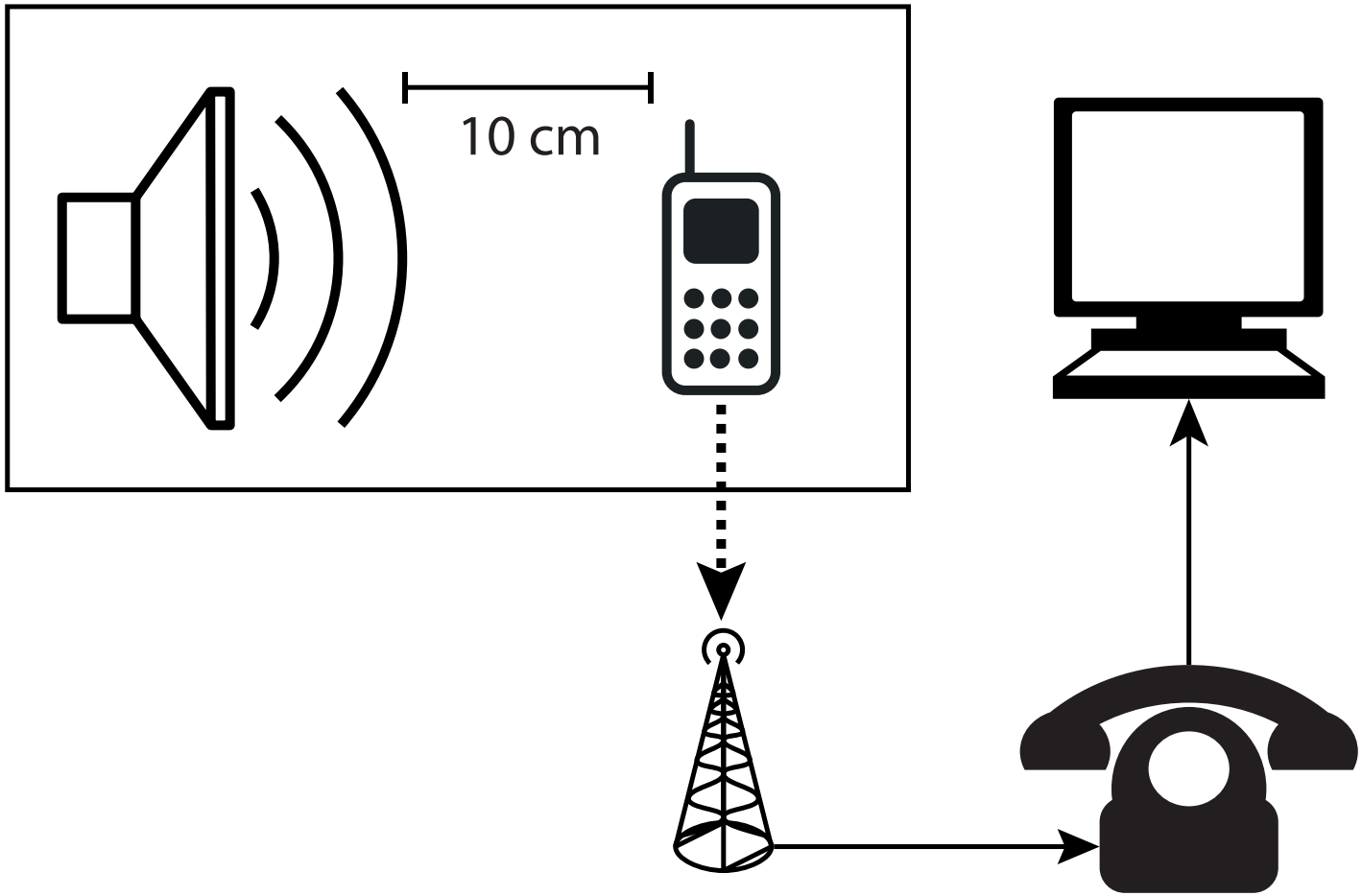
`http://databases.forensic-voice-comparison.net/`

Mobile-to-landline telephone transmission

- Microphone characteristic
- UMTS/GSM Adaptive Multi-Rate codec
  - Algebraic code-excited linear prediction (ACELP)
  - 8 similar modes with different bit rates
  - Link adaptation (mode can change every 40 ms (ETSI, 1999))
  - Discontinuous transmission / comfort noise generation

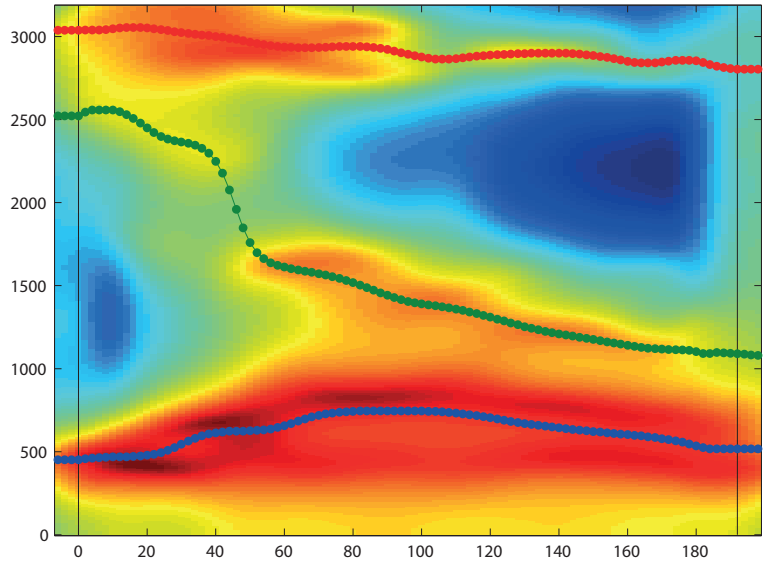
  (Other mobile telephone networks use EVRC-B (see e.g. Alzqhoul et al., 2012), AMR-WB, etc.)
- a-law compansion algorithm (ITU, 1988)
- Landline telephone bandpass

Simulation of mobile-to-landline transmission



Formant-trajectory system

- manually marked /iau/ tokens
- human-supervised formant-trajectory measurement

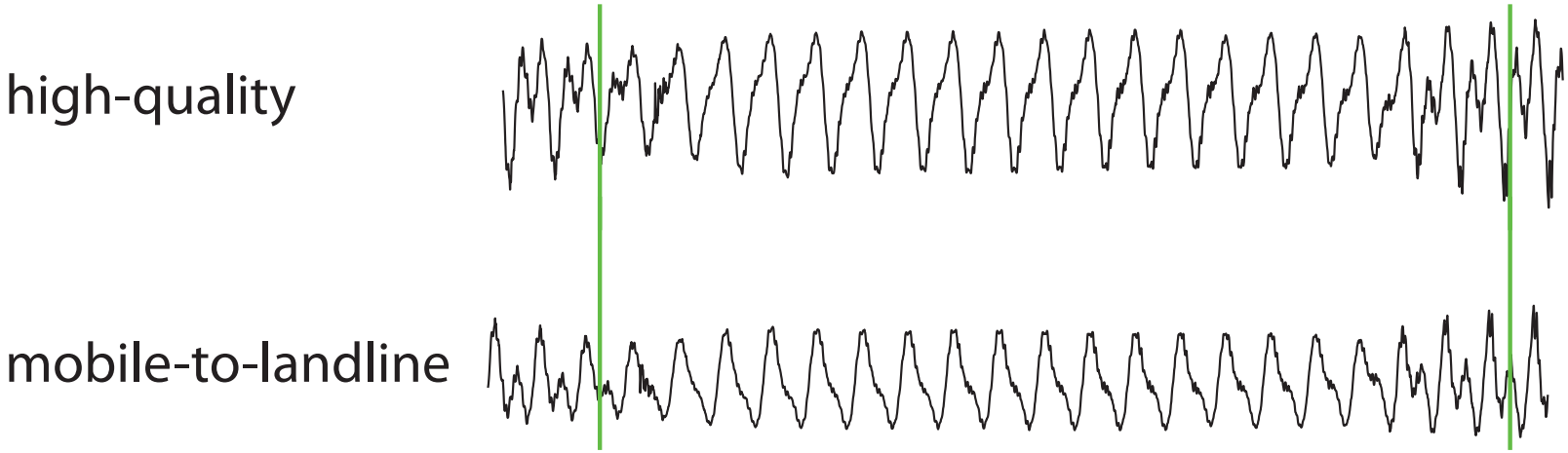


- 0th through 4th discrete cosine transform (DCT)
- coefficients of F2 and F3
- multivariate kernel density (MVKD) formula
- logistic-regression calibration using scores from development set

Compensation for recording-condition mismatch

Statistical compensation training using background data

- Parallel (aligned) high-quality and mobile-to-landline data
  ➡ Formant measurement from same speech segment

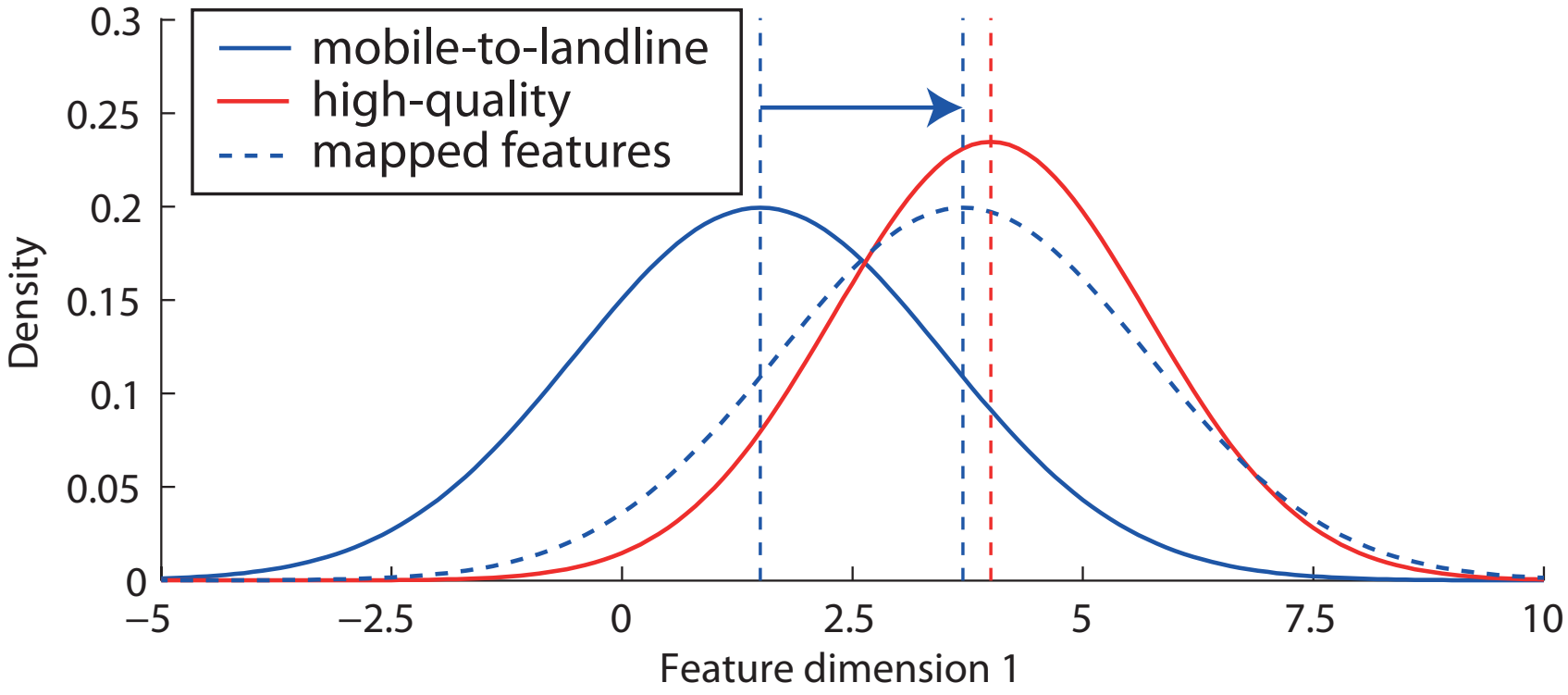


3 Methods:

1. Feature mapping
2. Canonical linear discriminant functions
3. Combining M1 and M2

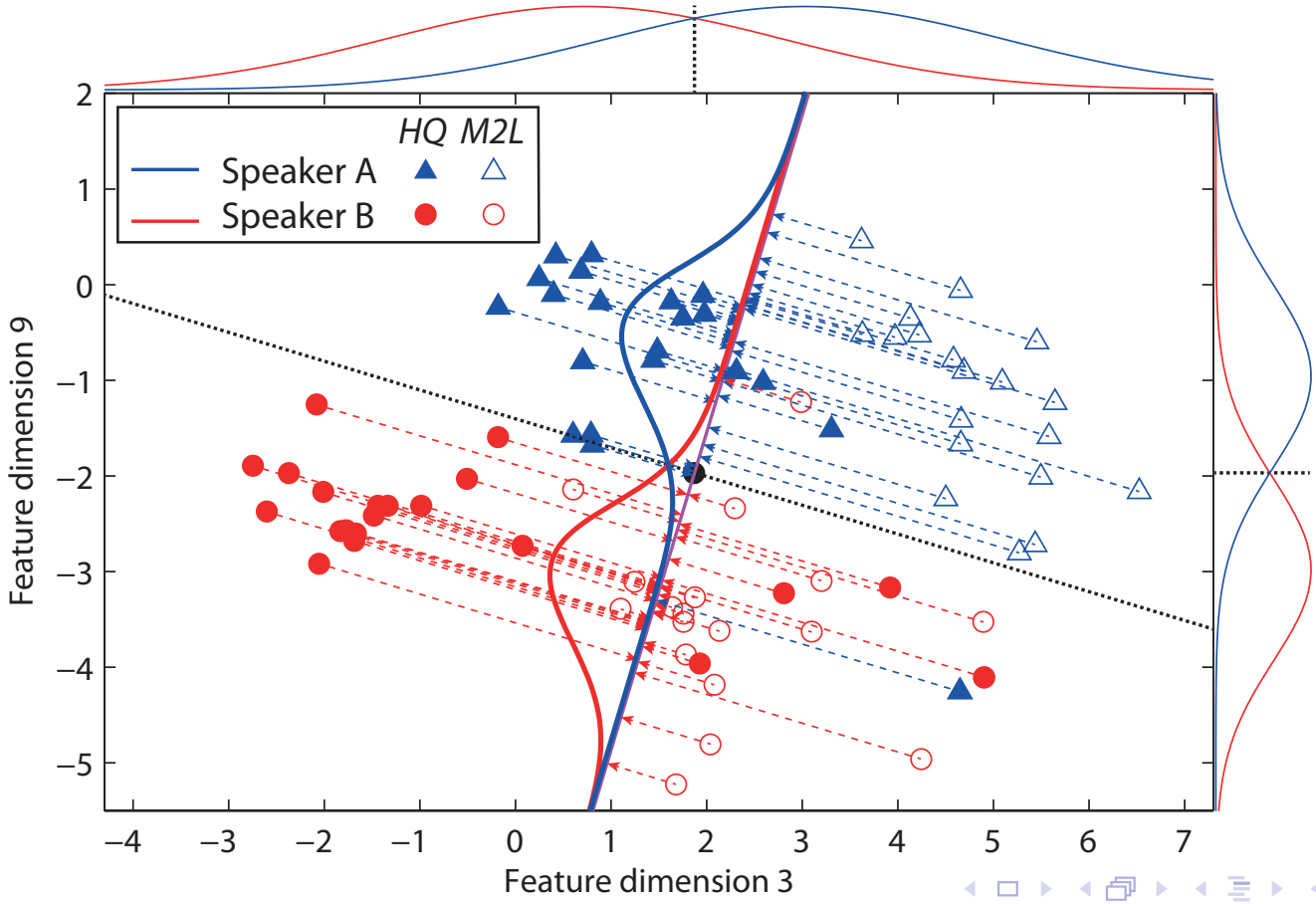
Feature mapping

- Calculate average difference between high-quality and mobile-to-landline DCT coefficients
- Use average of differences from multiple speakers as offset



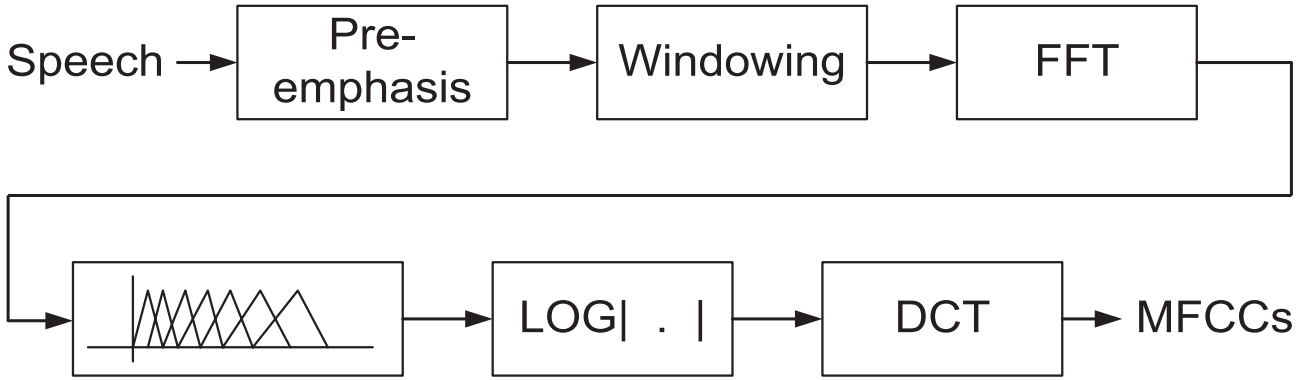
Canonical linear discriminant functions

- linear combinations of variables (DCTs) that are derived so that the groups in the training data are maximally separated on the new dimensions described by the functions
- both within- and between-group variation are taken into account



Fusion with 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 measures

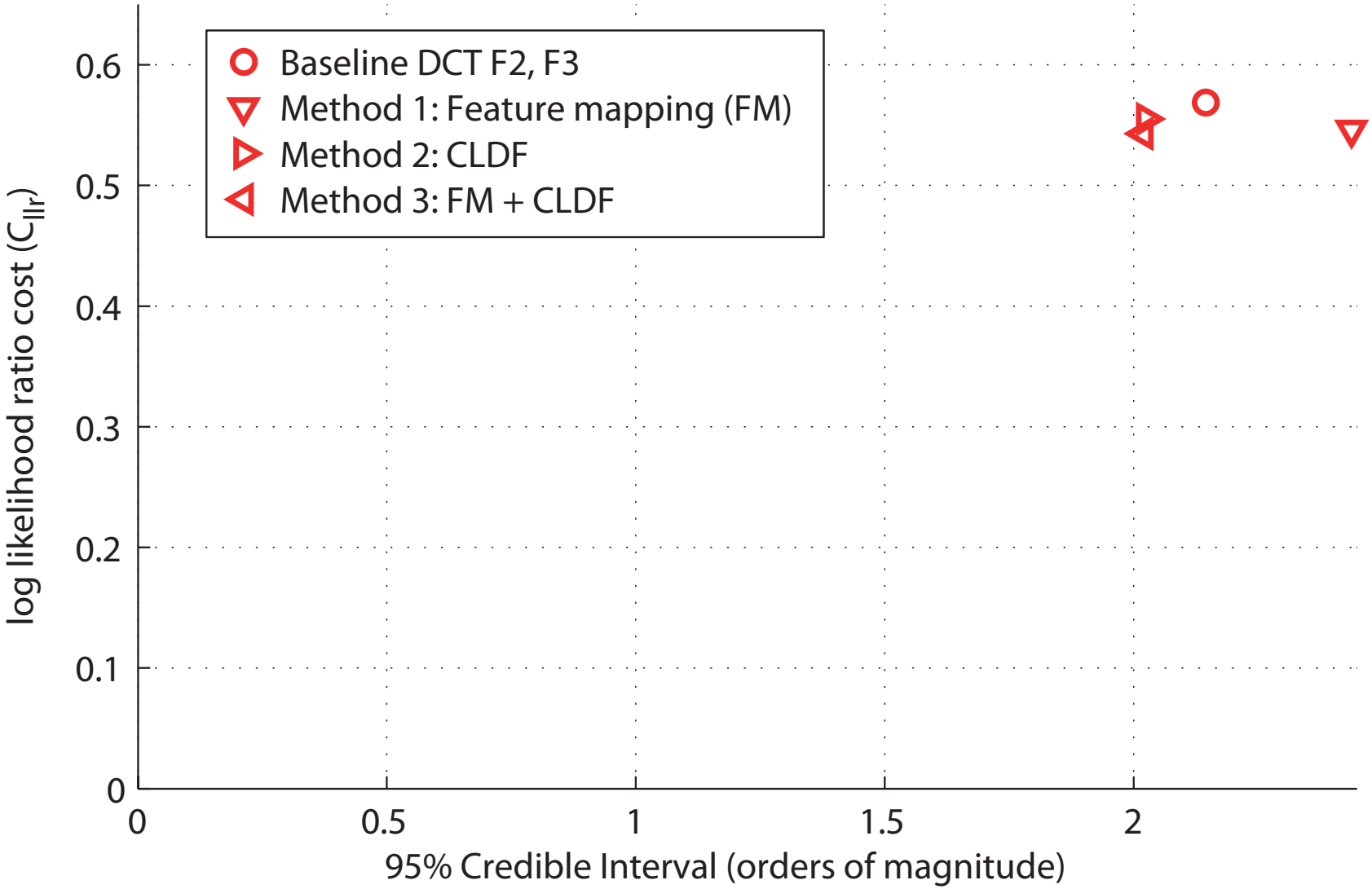
Validity / Accuracy:

- Log-likelihood ratio cost (C_{llr})

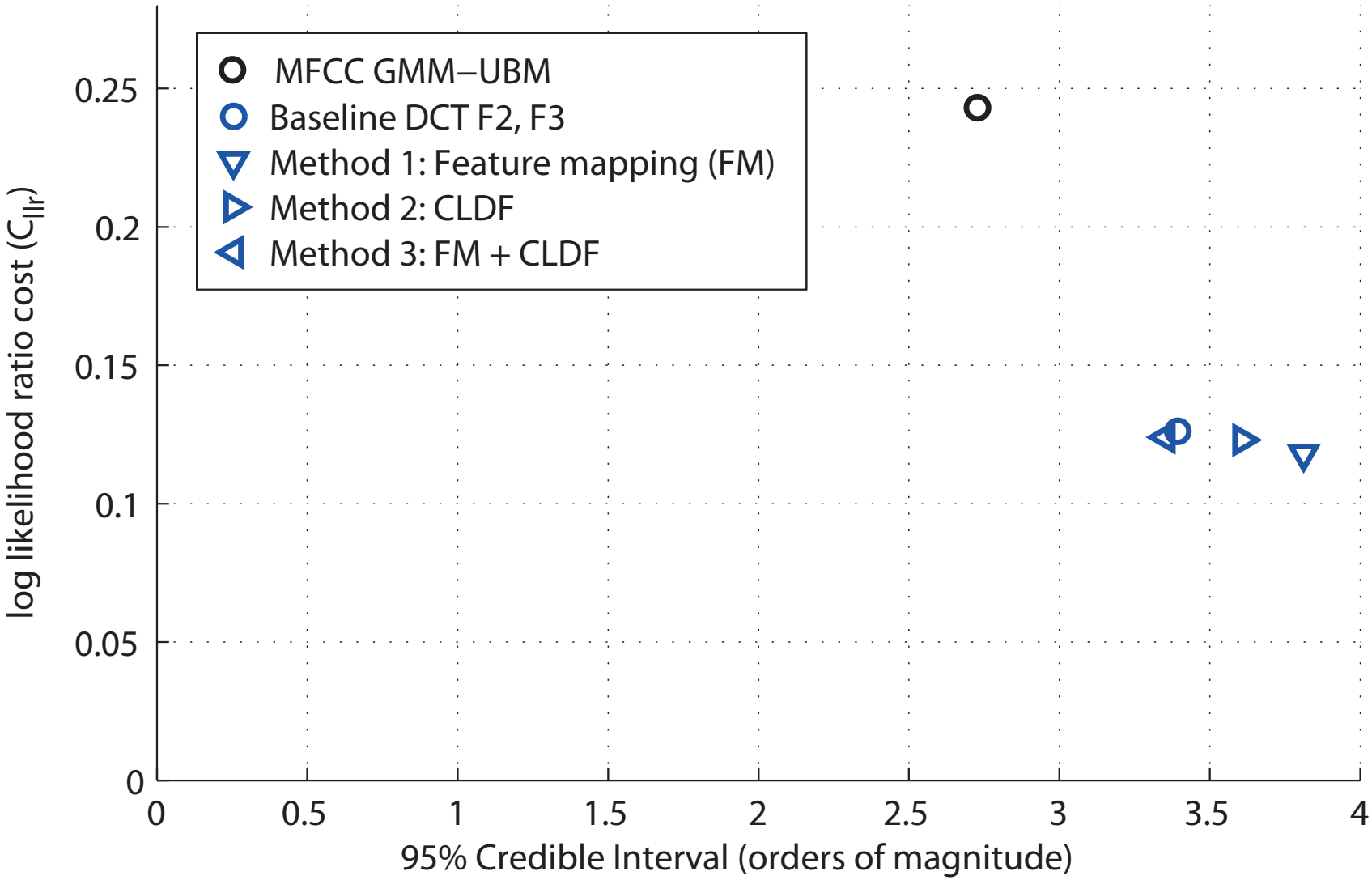
Reliability / Precision:

- 95% credible interval (Morrison, 2011)
- Parametric estimation method

Results before fusion



Results after fusion



Discussion & Conclusions

- Improvements in both validity and reliability were observed
- No substantial improvement after fusion with MFCC GMM-UBM system
- Potential reasons:
  - Differences in formant-trajectory measurements may cause non-linear effects in DCT coefficients

Thanks!

References

Alzqhoul, E. A. S., Nair, B. B. T., & Guillemin, B. J. (2012). Speech handling mechanisms of mobile phone networks and their potential impact on forensic voice analysis. In *Proc. SST-12*.

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *Int. J. of Speech, Lang. and the Law*, *11*(1), 83–102.

ETSI (1999). TS 101 709 Digital cellular telecommunications system (Phase 2+); Link Adaptation. `http://www.etsi.org/deliver/etsi_ts/101700_101799/101709/08.01.00_60/ts_101709v080100p.pdf`

ITU (1988). ITU-T Recommendation G.711 (11/88): Pulse code modulation (PCM) of voice frequencies.

Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurment of formant frequencies. *Forensic Linguistics*, *8*(1), 80–99.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Sci. Justice*, *51*, 91–98.

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 Commun.*, *55*, 796–813.