The effect of session variability on the validity and reliability of forensic-voice-comparison systems

Ewald Enzinger

Geoffrey Stewart Morrison

FORENSIC VOICE COMPARISON LABORATORY
SCHOOL OF ELECTRICAL ENGINEERING & TELECOMMUNICATIONS

UNSW
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

Introduction

- There is always a time lag between recording of offender and suspect
  - between session variability
  - however, researchers sometimes test their systems using within-session data for same-speaker comparisons
  - How does this compare with between-session testing?

Data

- 60 female Standard Chinese speakers
- Two recording sessions separated by 2–3 weeks
- Channels:
  - high quality
  - mobile-to-landline
- Split into 3 groups of 20 speakers
  - background database
  - development set
  - test set

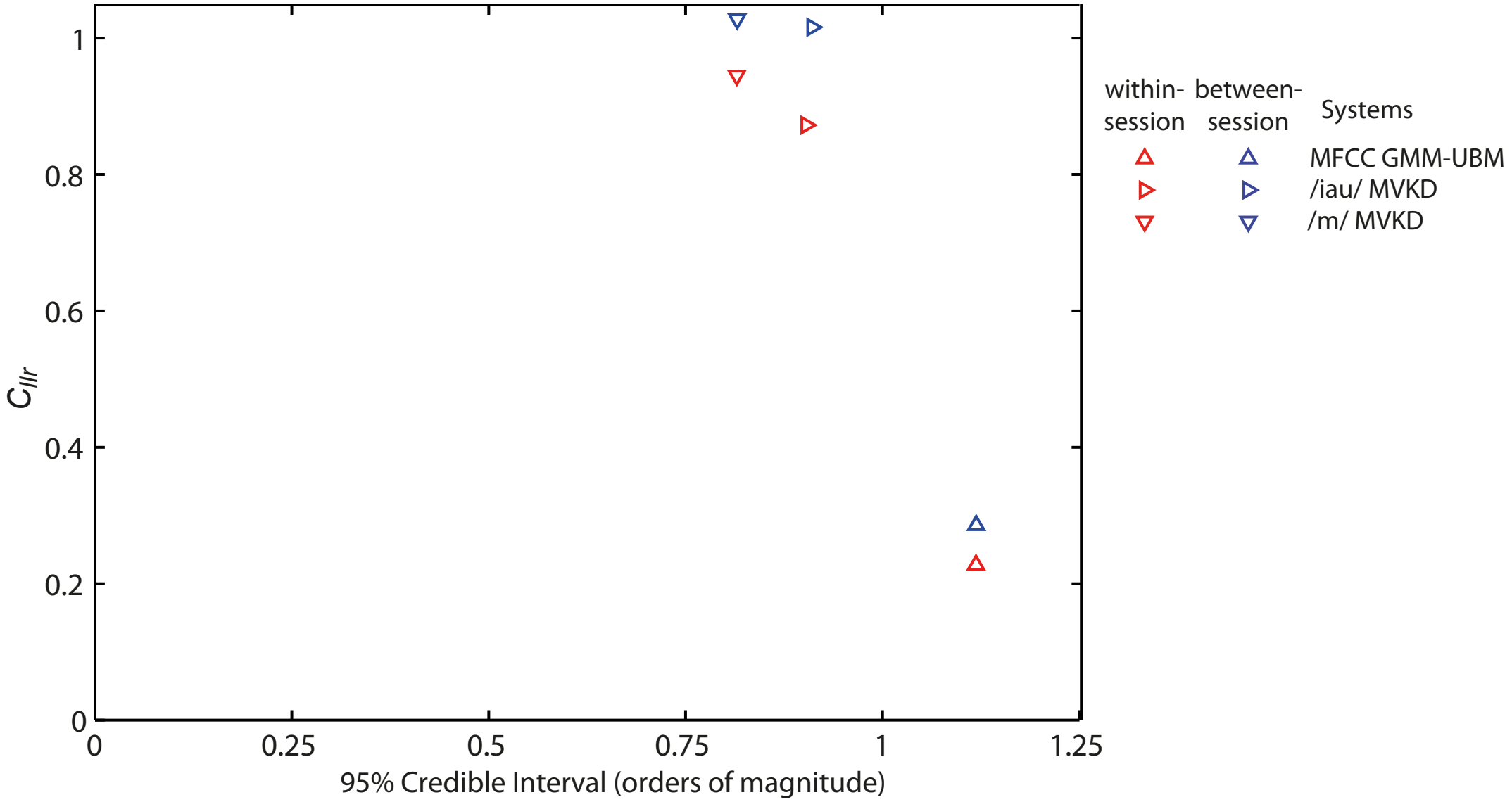
Forensic-voice-comparison systems

- 3 systems:
  - formant-trajectories of /iau/ tokens, MVKD
  - cepstral coefficients of /m/ tokens, MVKD
  - MFCCs + deltas, GMM-UBM
- Logistic-regression calibration

Procedures

- Same amount of data used for within-session and between-session tests
- Channels used:
  - mobile-to-landline for offender recording
  - high-quality for suspect recording
  - high-quality for background recordings

Results


1
0.8
0.6
0.4
0.2
0
C_{llr}
0
0.25
0.5
0.75
1
1.25
95% Credible Interval (orders of magnitude)
within-session
between-session
Systems
MFCC GMM-UBM
/iau/ MVKD
/m/ MVKD

Conclusion

- In casework the offender and suspect samples are always between session
- Testing validity and reliability on within-session data gives overly optimistic results
- Validity and reliability must be tested using between-session data