Special Session on Forensic Acoustics
162nd Meeting of the Acoustical Society of America
San Diego, 2 November 2011
http://sandiego2011.forensic-acoustics.net/

Comparison of human-supervised and fully-automatic formant-trajectory measurement for forensic voice comparison

Cuiling Zhang
Felipe Ochoa
Ewald Enzinger
Geoffrey Stewart Morrison

FORENSIC VOICE COMPARISON LABORATORY
SCHOOL OF ELECTRICAL ENGINEERING & TELECOMMUNICATIONS

UNSW
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

Acknowledgment of Funding

- Data collection was funded by an International Association of Forensic Phonetics and Acoustics (IAFPA) Research Grant.

- Analysis was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

- Presentation supported by the Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142. Unless otherwise explicitly attributed, the opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations.

Research Questions

- Formant measurement is very common in acoustic-phonetic forensic voice comparison.

- What is the reliability of human supervised formant-trajectory measurement?

- What is the validity and reliability of forensic-voice-comparison systems based on:
  - human-supervised formant-trajectory measurement?
  - fully-automatic formant-trajectory measurement?

- Improvement over baseline fully-automatic MFCC system?

Data

- 60 female speakers of Standard Chinese
  - 20 for background
  - 20 for development
  - 20 for test
- Information-exchange task over the telephone
- Two recording sessions separated by 2–3 weeks
  - 4–5 minutes of speech per speaker per session
- High-quality audio
- Chinese /iau/ tokens
  - 15–30 tokens per speaker per recording

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

Acoustic Analysis

- Manual marking of /iau/ tokens (CZ)
  - SoundLabeller
- Human-supervised formant tracking
  - FormantMeasurer
  - 4 supervisors (CZ, EE, FO, GSM)
  - 3 blocked measurement repetitions per supervisor on session 1
  - 1 measurement by CZ and FO on session 2
- Fully-automatic formant tracking
  - WaveSurfer
  - Praat
  - Nearey, Assmann, Hillenbrand (2002) [NAH2002]
  - Mustafa, Bruce (2006) [MB2006]
  - Rudoy, Spendley, Wolfe (2007) [RSW2007]

Forensic-voice-comparison systems

- Discrete cosine transform (DCT)
  - F2 and F3
  - zeroth through fourth coefficients
- Multivariate kernel density (MVKD) formula
- Logistic-regression calibration
- Baseline mel-frequency cepstral coefficient (MFCC) Gaussian-mixture model universal-background model (GMM-UBM)
  - Entire speech-active portion of recording
- Logistic-regression fusion

Results & Discussion

- Reliability of human-supervised formant measurement
  - within-supervisor standard deviation
  - across all formants, all tokens, all speakers

	σ	
supervisor	Hz	%
CZ	49	2.4
EE	59	2.8
FO	58	3.0
GSM	50	2.4
between	77	3.8

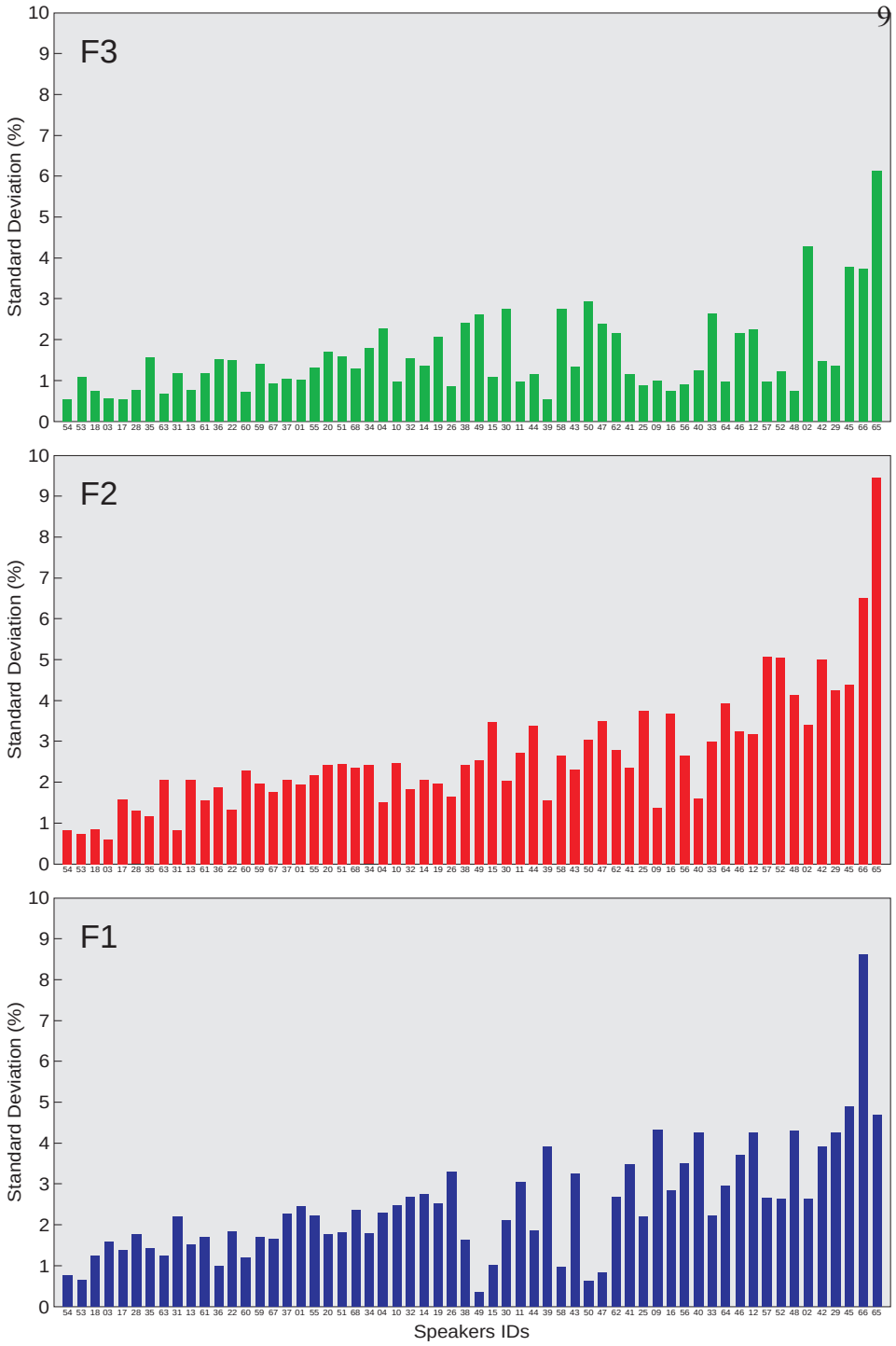
Results & Discussion

- Reliability of human-supervised formant measurement
  - within-supervisor standard deviation
  - across all tokens, all speakers

	σ					
	Hz			%		
supervisor	F1	F2	F3	F1	F2	F3
CZ	20	51	66	2.9	2.9	1.9
EE	21	65	77	2.8	3.4	2.1
FO	23	54	81	3.4	3.3	2.1
GSM	17	47	71	2.5	2.7	1.8
between	26	79	105	4.0	4.2	2.9

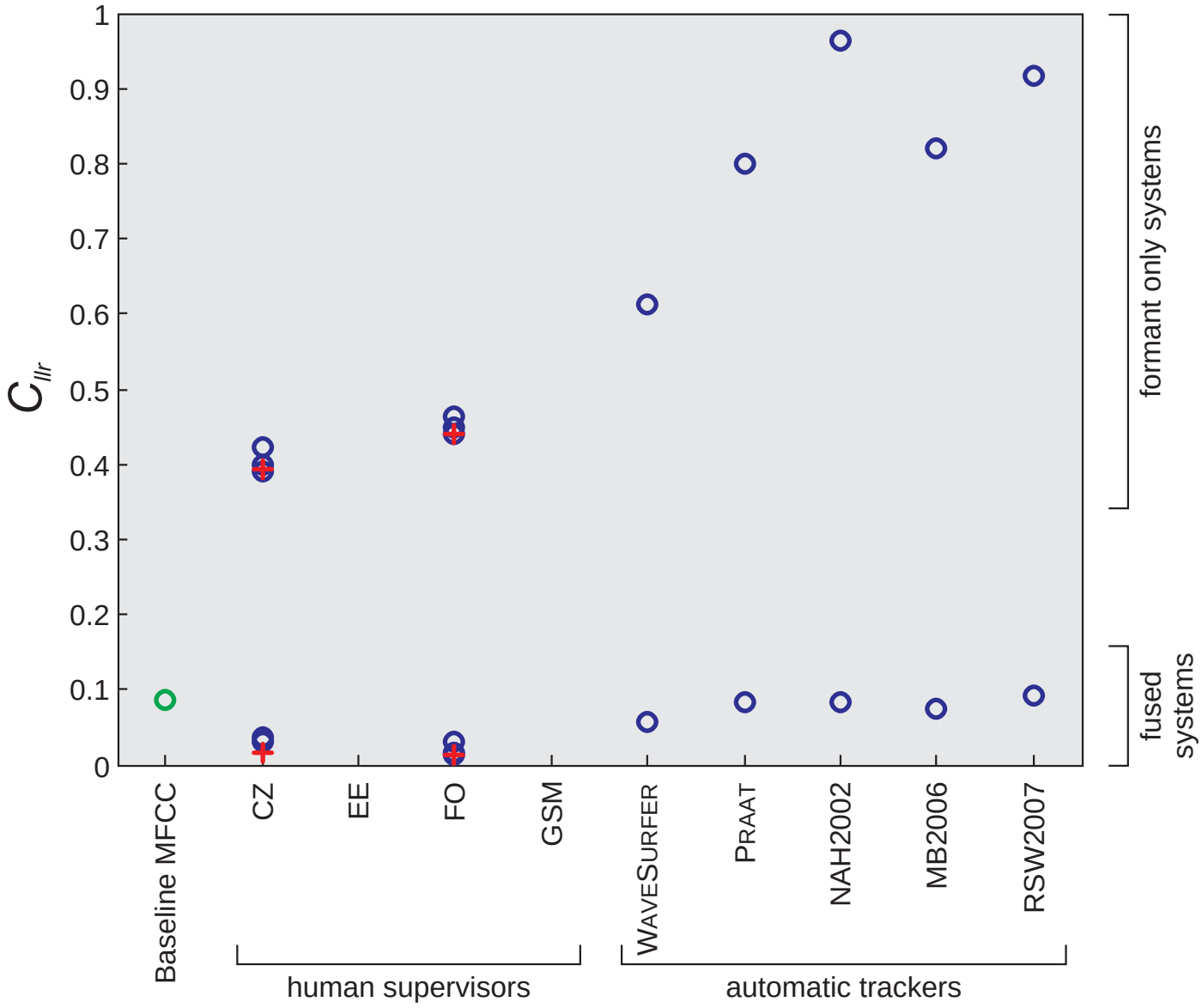
Results & Discussion

- Reliability of human-supervised formant measurement
  - within-speaker standard deviation
  - across all tokens, all supervisors



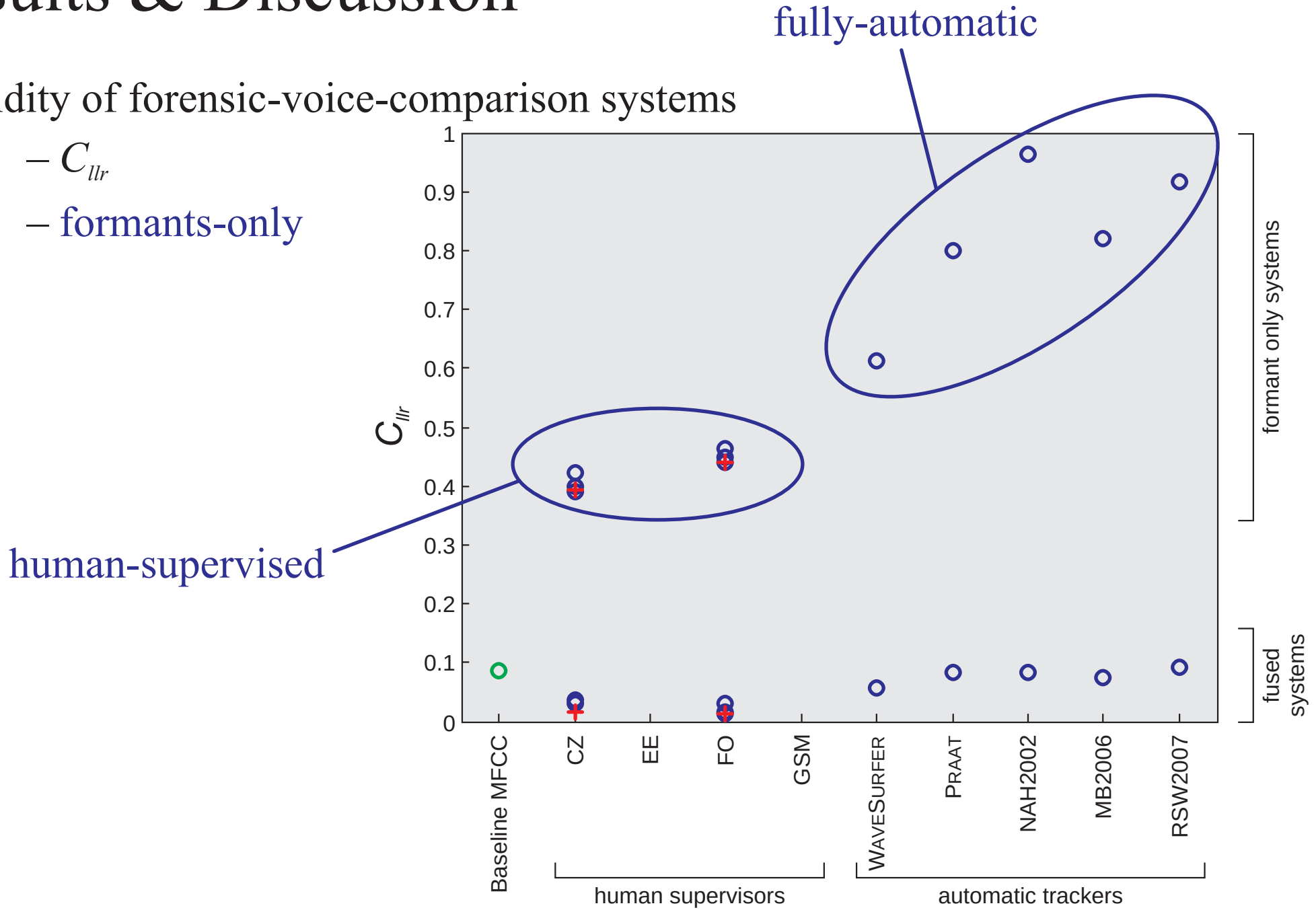
Results & Discussion

- Validity of forensic-voice-comparison systems
  - C_{llr}



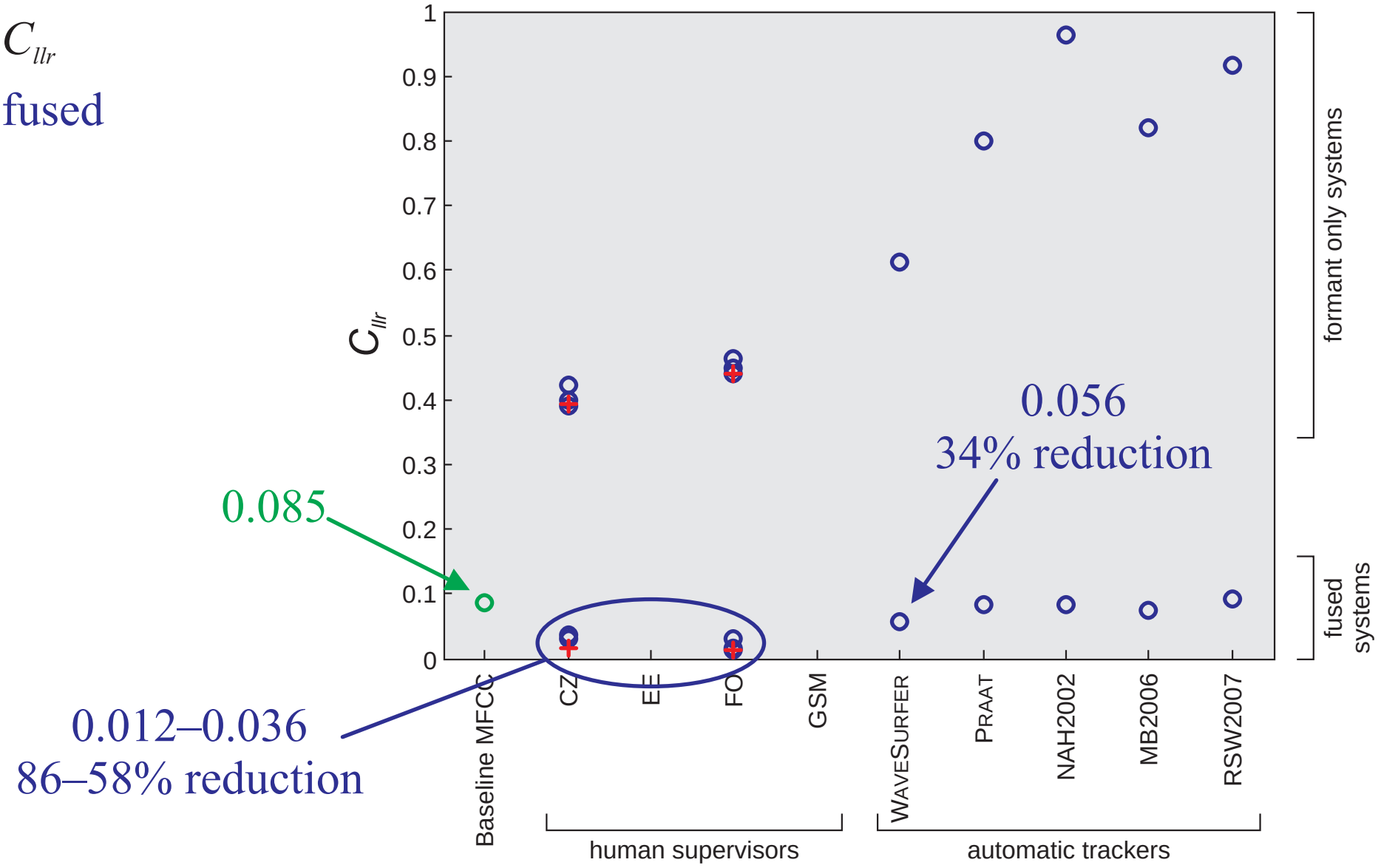
Results & Discussion

- Validity of forensic-voice-comparison systems
  - C_{llr}
  - formants-only



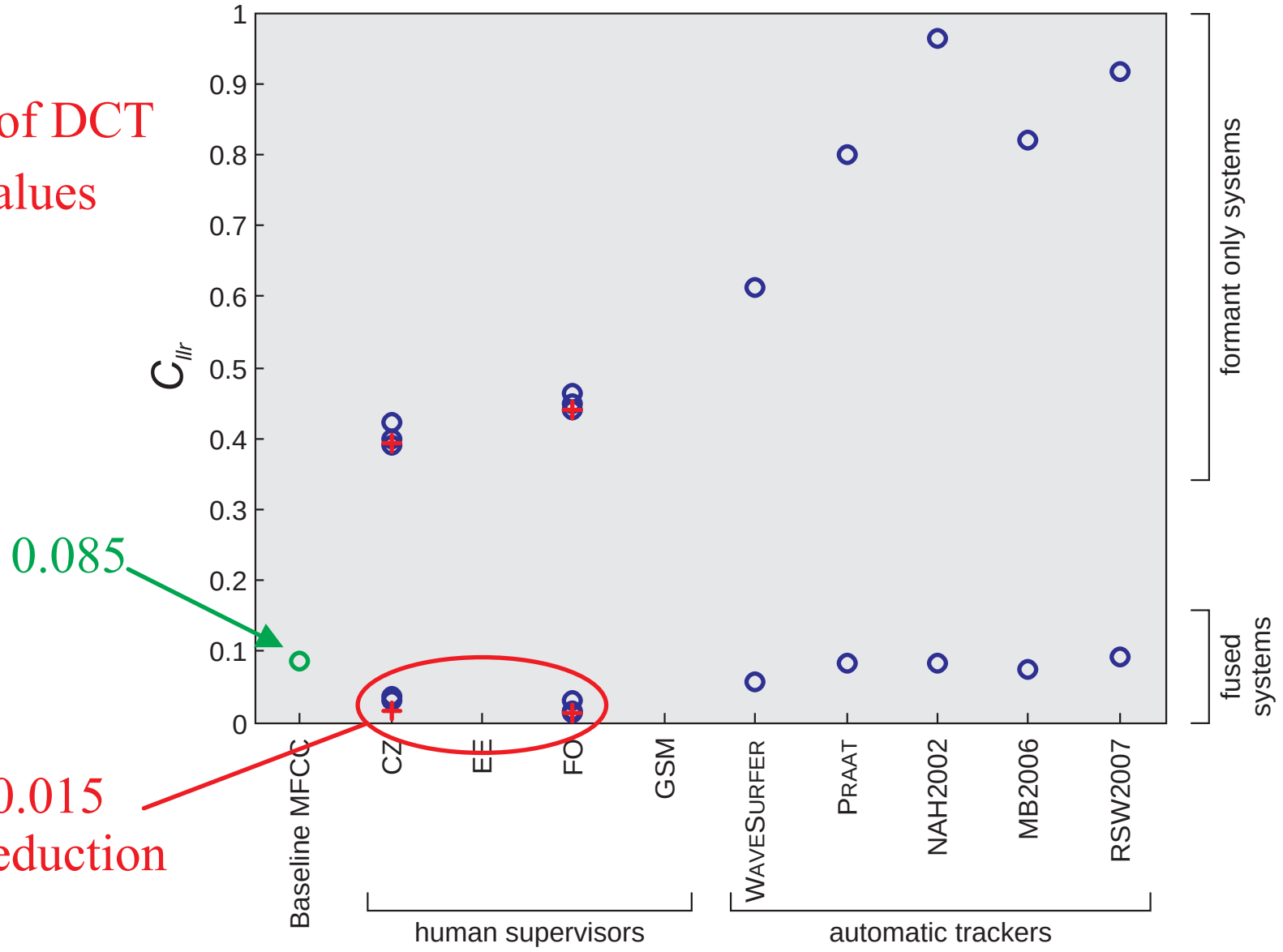
Results & Discussion

- Validity of forensic-voice-comparison systems
  - C_{llr}
  - fused



Results & Discussion

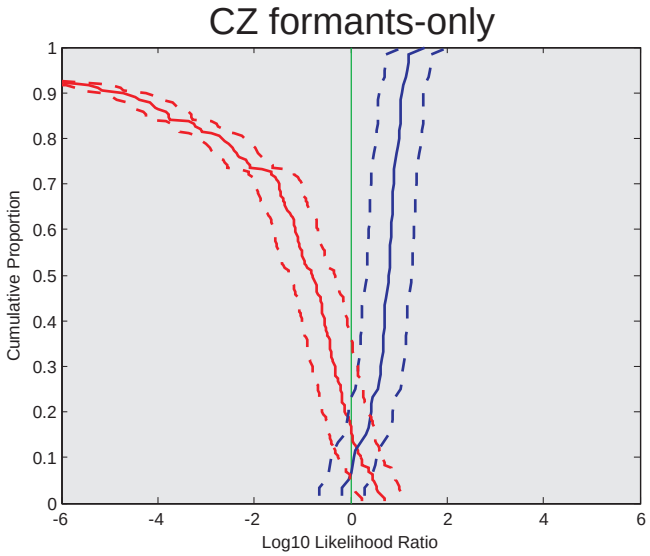
- Validity of forensic-voice-comparison systems
  - C_{llr}
  - central set of DCT coefficient values per formant per token



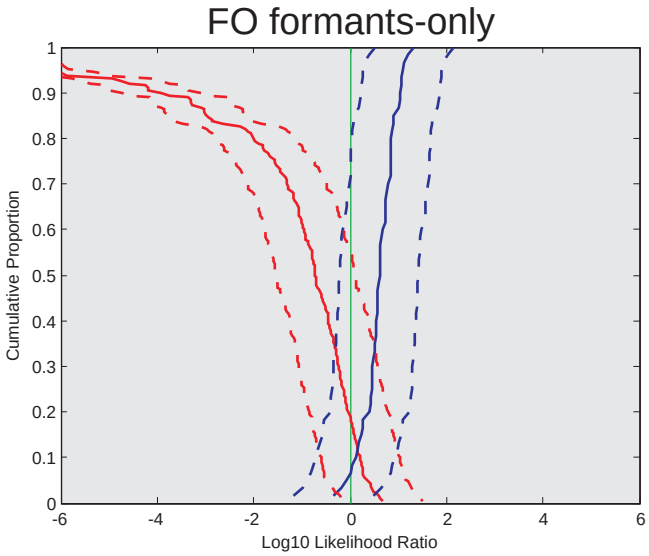
Results & Discussion

- Reliability of forensic-voice-comparison systems based on human-supervised formant-trajectory measurments
  – 95% credible interval in $\log_{10}$ LR

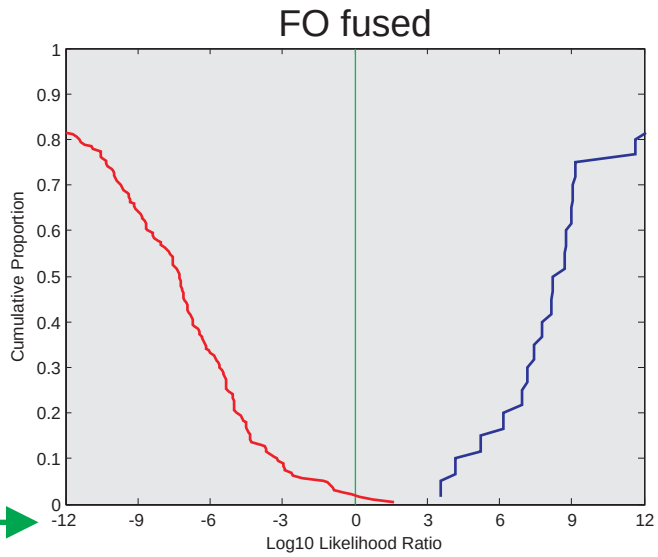
±0.47



±0.82



±0.00



different scale →

Conclusions

- Fused systems incorporating human-supervised formant-trajectory measurements: Major improvements over baseline MFCC system
- Fused systems incorporating central measurements from human-supervised formant-trajectory measurements: Best performance
- Fused system incorporating fully-automatic WAVESURFER formant-trajectory measurements: Substantial improvement over baseline MFCC, but not as good as systems using human-supervised formant measurements
- Is the reliability of human-supervised formant-trajectory measurement acceptable for forensic voice comparison?
- Are the human-labor costs warranted?

Thank You