The complementarity of automatic, semi-automatic and phonetic measures of vocal tract output

Vincent Hughes, Philip Harrison, Paul Foulkes,
Peter French, Colleen Kavanagh & Eugenia San Segundo

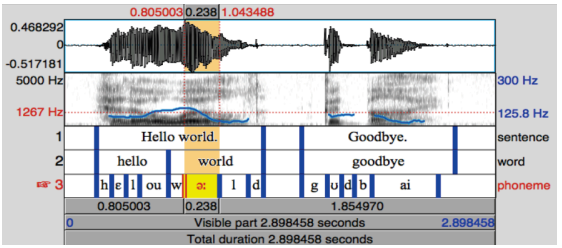
J P French λssociates
Forensic speech and acoustics laboratory

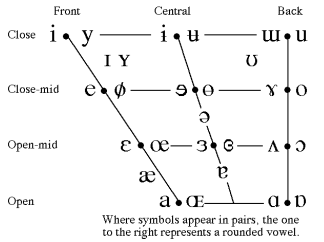
IAFPA
9-12 July 2017

1. Forensic voice comparison (FVC)

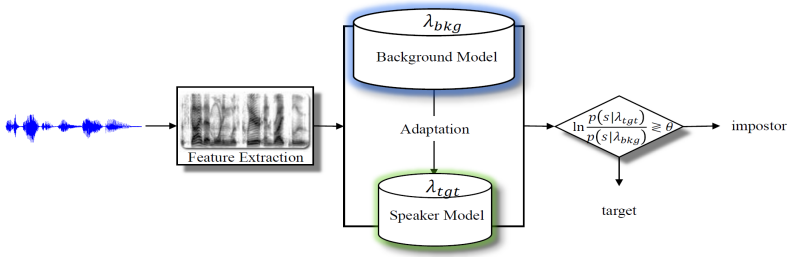
- three common methods of analysis:

linguistic-phonetic

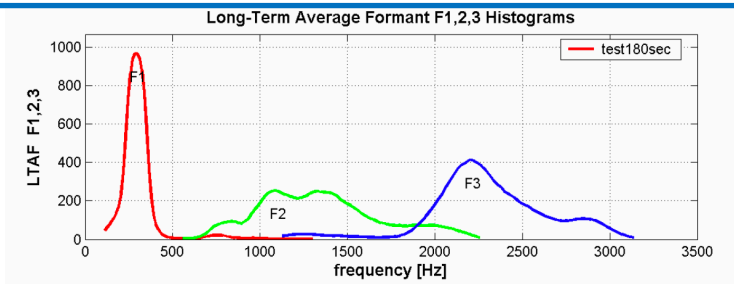




automatic (ASR)



semi-automatic (S-ASR)



1. FVC: Combining approaches

- largely developed in isolation
  - but ultimate aim is the same…
- increasing focus on combination of (S-)ASR and ling-phon approaches
  - (H)ASR element of NIST (Greenberg et al. 2010)
  - G'ment labs in Germany and Sweden use combined approach in casework
  - Zhang et al (2013), Gonzalez-Rodriguez et al (2014)

2. This study: Features

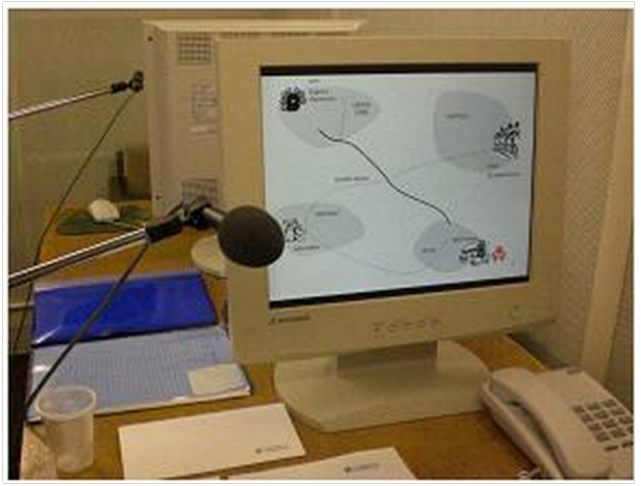
- measures of long term vocal tract (VT) output
  - **automatic:** MFCCs
  - **semi-automatic:** LTFDs
  - **ling-phon:** supralaryngeal voice quality (VQ)
- why?
  - commonly used in each approach
  - encode considerable speaker information
  - in principle model the same thing

2. This study: Research questions

1. how does the performance of MFCCs and LTFDs compare on the same data?
2. does fusion of MFCC and LTFD systems improve performance over MFCCs only?
3. can supralaryngeal VQ explain the *errors* made by the (S-)ASR system?

With a view to the future… what about laryngeal VQ?

3. Method



- DyViS (Nolan et al. 2009)
  - **Task 1:** mock police interview
  - **Task 2:** telephone conversation with accomplice

- pre-processing
  - manual editing
  - silences (> 100ms) removed
  - sections of clipping removed

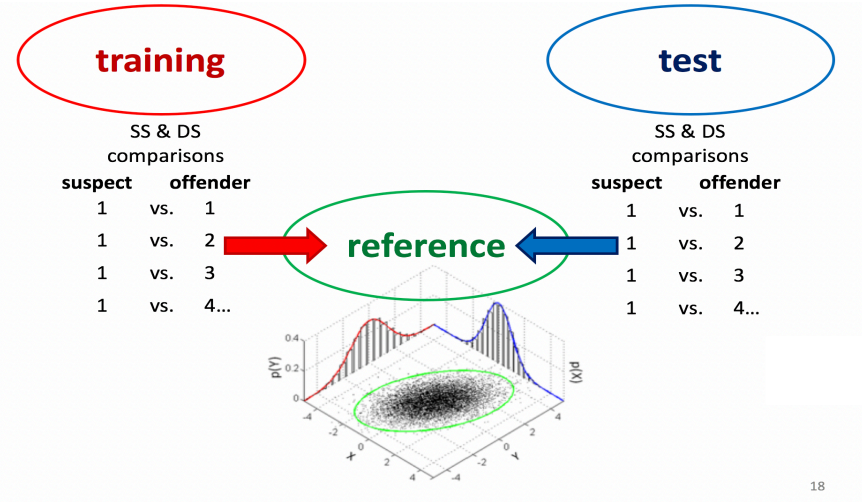
3. Method: (S-)ASR

- for (**semi-**)**automatic** features:
  - audio segmented into Cs and Vs (StkCV)
  - **94**/100 speakers with > 60s of Vs
  - samples reduced to 60s net Vs (6000 frames)
  - 20ms frames/ 10ms shift (hamming window)

MFCCs	**LFTDs**	**(M)LTFDs**
12 MFCCs	F1-F4 frequencies	F1-F4 (Mel) frequencies
12 Δs	F1-F4 Δs	F1-F4 (Mel) Δs
12 ΔΔs	F1-F4 bandwidths	F1-F4 (Mel) bandwidths

3. Method: (S-)ASR

- 94 speakers divided into sets:
  - training (31 speakers)
  - test (31 speakers)
  - reference (32 speakers)



- SS and DS LRs computed
  - Task 1 = suspect/ Task 2 = offender
  - GMM-UBM (w. MAP adaptation)

3. Method: (S-)ASR

- logistic regression calibration/fusion:
  - applied separately for individual and combined systems

- system validity:
  - Equal error rate (EER):
  - Log LR Cost Function (C_{llr}; Brümmer & du Preez 2006)

3. Method: Voice quality

- auditory based analysis using modified VPA
  - Laver et al (1981); San Segundo et al (submitted)
  - 25 supralaryngeal features
  - 7 laryngeal features

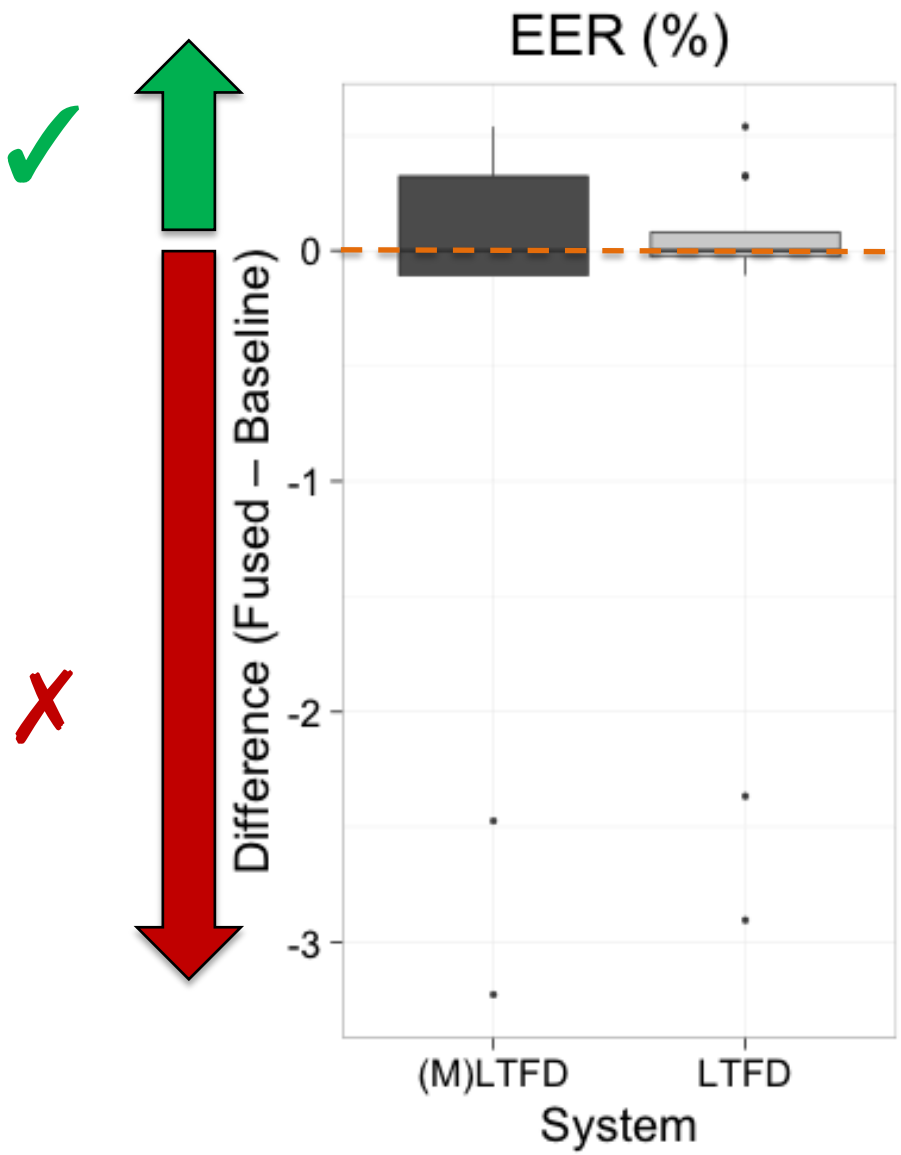
- Task 2: 100 speakers
  - PFo, PFr, ESS produced VPAs independently
  - agreed VPA profiles (after calibration)

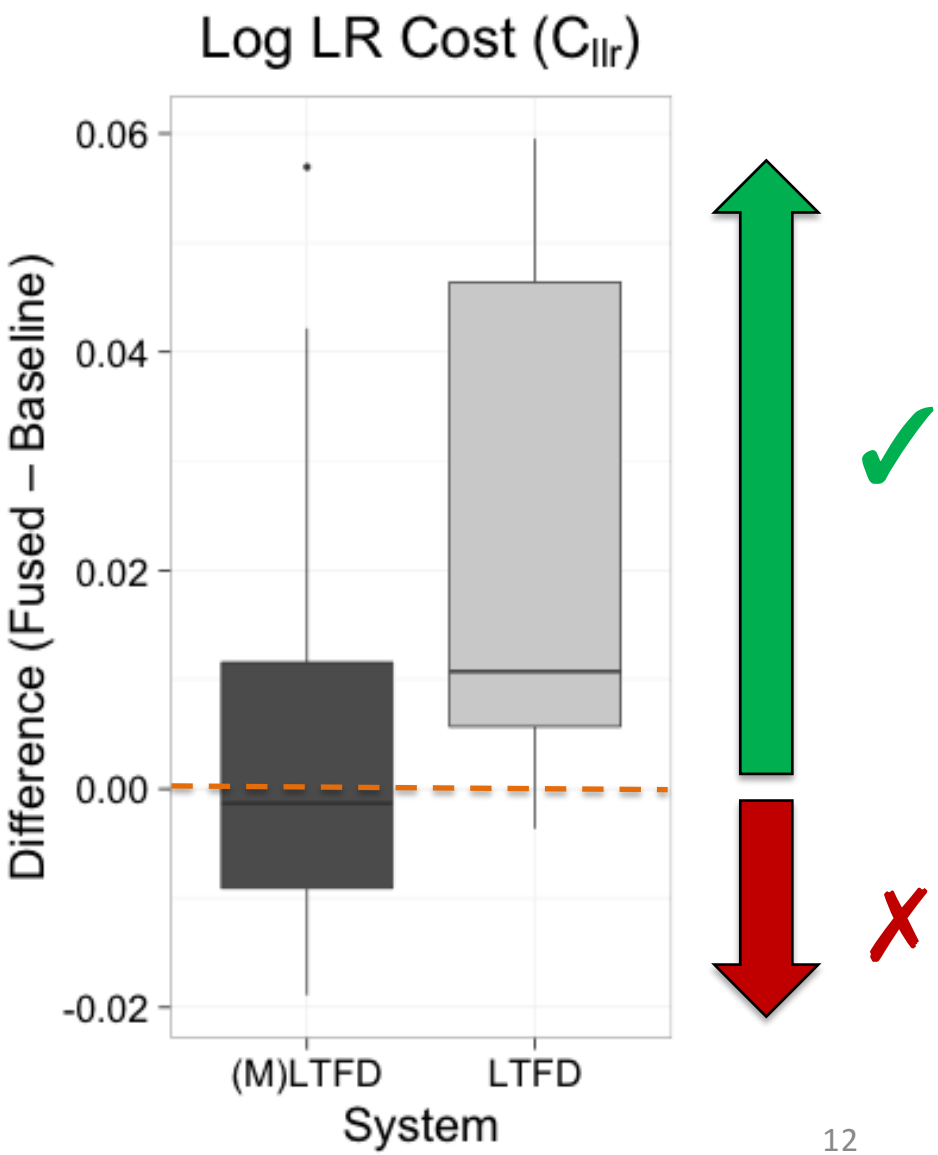
4. Results: MFCCs and LTFDs

System	EER (%)	C_{llr}
MFCCs	6.45	0.267
MFCCs+Δs	3.55	0.190
MFCCs+Δs+ΔΔs	**3.23**	**0.146**

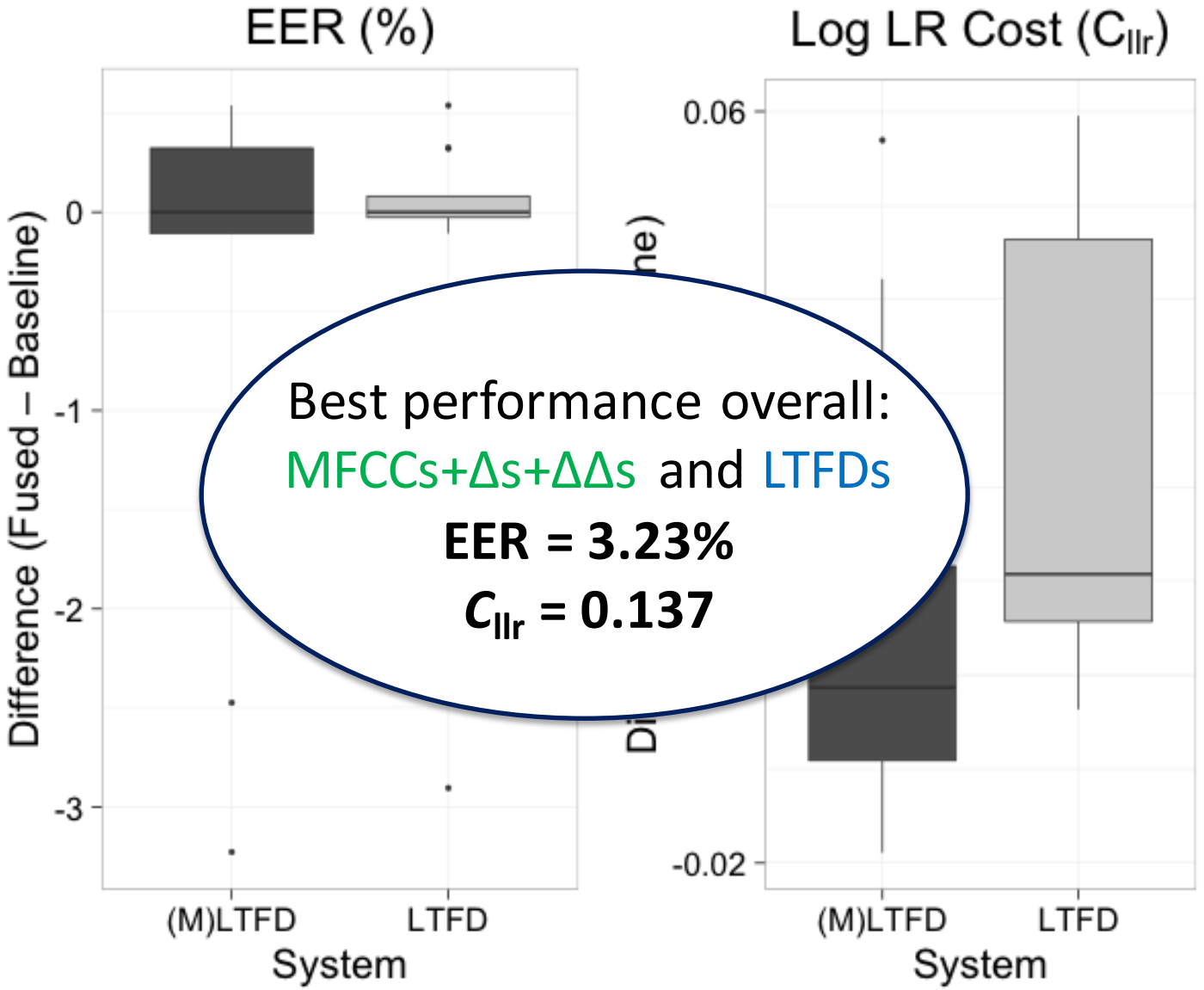
System	EER (%)	C_{llr}
LTFDs	6.67	0.284
LTFDs+BWs	**6.45**	**0.255**
LTFDs+ Δs	6.99	0.311
LTFDs+BWs+Δs	6.67	0.259
(M)LTFDs	**8.29**	**0.290**
(M)LTFDs+BWs	9.68	0.462
(M)LTFDs+Δs	9.57	0.325
(M)LTFDs+BWs+Δs	9.68	0.507

4. Results: Fusion





4. Results: Fusion



4. Results: Supralaryngeal VQ

- best system = 14 *errors*
  - 13 false acceptances (DS producing SS evidence)
  - what is it about these speakers?
- 9 involved speakers **#067** and **#072**
  - fairly typical supralaryngeal VQ profiles
  - non-neutral for:

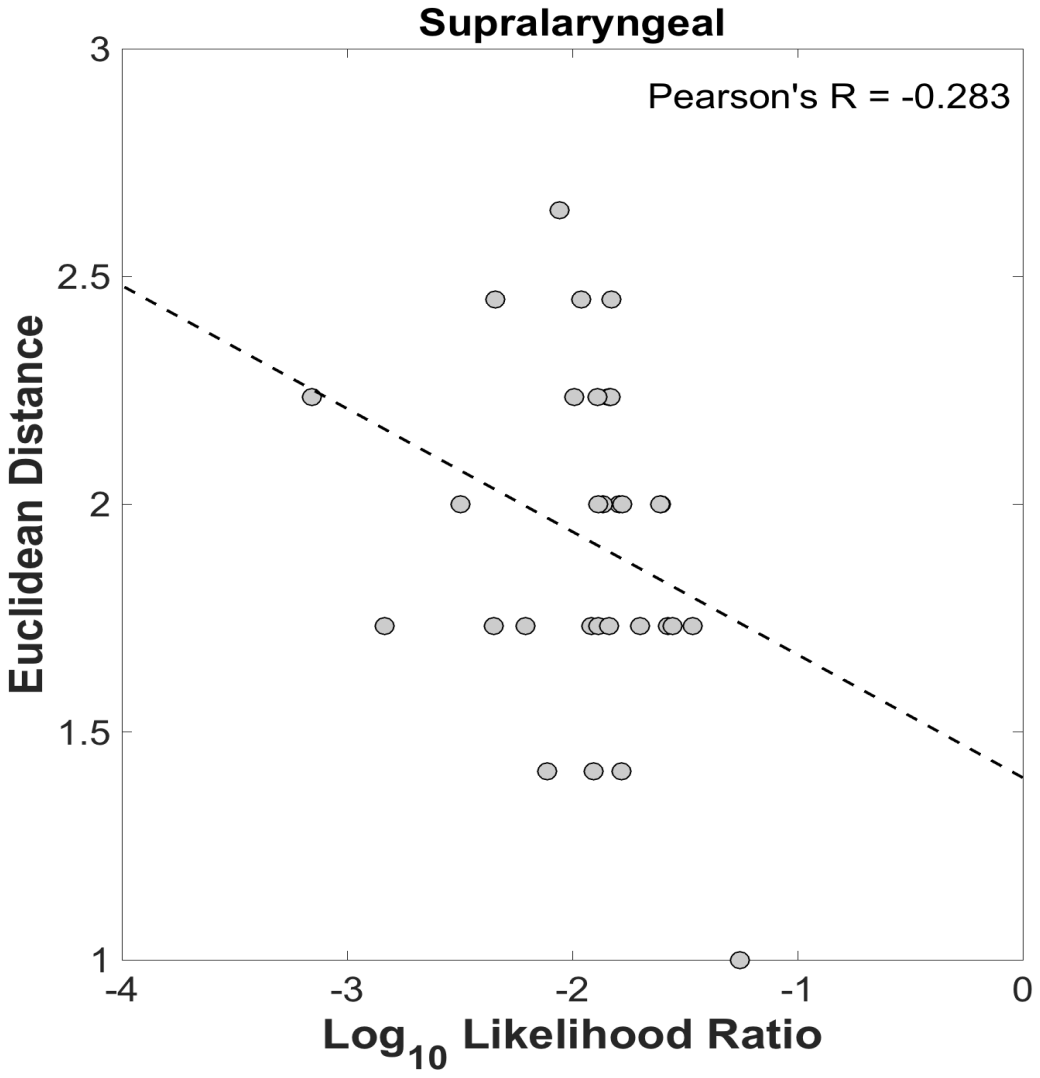
    advance tongue tip, fronted tongue body, nasality
  - easily confused with other speakers?

4. Results: Supralaryngeal VQ

atypical

VQ

typical



less confusable ← (S-)ASR → more confusable

5. Discussion

- MFCCs outperform LTFDs and (M)LTFDs
  - Mel weighting of LTFDs = worse
- fusion of formants and MFCCs: **no improvement** in performance
  - MFCCs encode the same speaker-discriminatory information as formants
  - MFCCs = richer representation/ higher resolution
    - capture more speaker information

5. Discussion

- errors produced by (S-)ASR = explainable using supralaryngeal VQ
  - speakers with generic supralaryngeal VQ profiles are more difficult for the (S-)ASR system to separate
- trend = weak, but impressive given…
  - ASR based only on vowels/ VQ on all data
  - VQ = auditory-based, relatively blunt tool
  - MFCCs = mathematically abstract, rich in information
  - averaging over all DS LLRs & all VPA features

5. Discussion

- so… can we resolve the errors?
  - 14 *error* pairs presented to two experts blind
  - instructed to use auditory analysis only and make decisions relatively quickly
  - outcome = LR-like scores
- both experts correctly classified all pairs
  - task = relatively straightforward
  - relied primarily on laryngeal VQ

6. Conclusions

- evaluation of complementarity of different measures of VT output
- more work needed at the intersection of ASR and ling-phon FVC
  - important not to see methods as opposed
  - tools in the toolkit
- future: potentially considerable value in looking at laryngeal VQ

Thanks!

Questions?

Special thanks to:
Richard Rhodes, Jessica Wormald, George Brown,
Jonas Lindh, Frantz Clermont

UNIVERSITY of York

J P French λssociates
Forensic speech and acoustics laboratory

IAFPA
9-12 July 2017

Arts & Humanities Research Council

voice and identity
0101101ID1