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

University of York
J P French Associates
Forensic speech and acoustics laboratory

Vincent Hughes, Philip Harrison, Paul Foulkes, Peter French
Colleen Kavanagh, Eugenia San Segundo
{vincent.hughes|philip.harrison|paul.foulkes|peter.french}@york.ac.uk

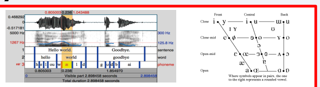
Arts & Humanities Research Council
voice and identity

1. Introduction

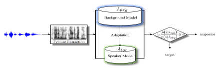
Forensic voice comparison (FVC) = offender (unknown) vs. suspect (known)

Three common methods of analysis

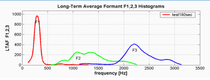
linguistic-phonetic



automatic (ASR)



semi-automatic (SASR)



- increasing focus on integrating methods in **research** ((H)ASR: Greenberg et al 2010; Hughes, Foulkes and Wood 2016) and **casework** (BKA Germany, Sweden)

Fundamental issues

- strengths and weaknesses of different methods
- do different methods capture the same or different speaker-specific information?
- *front-end* prediction of problem speakers for ASRs (*black box* perception in the Courts; see R v Slade and Ors [2015])
- improvement in FVC system performance using combined methods

Features for analysis

Voice quality (VQ)

- quasi-permanent vocal settings (supralaryngeal and laryngeal)
- regularly analysed in casework (Gold and French 2011)

Mel frequency cepstral coefficients (MFCCs):

- rich representation of the Mel-weighted power spectrum, decoupling supralaryngeal from laryngeal information

Long term formant distributions (LTFDs)

- formant values extracted automatically from all vocoids in speech stream

- most commonly used features in each FVC method
- known to encode considerable speaker-specific information
- all, in principle, capture information about the supralaryngeal vocal tract

2. Research questions

1. how does the performance of MFCCs and LTFDs compare?
2. does fusing MFCC + LTFD systems improve performance of MFCC only?
2. can supralaryngeal VQ explain the *errors* made by the (S-)ASR system?
3. what is the potential value of laryngeal VQ to (S-)ASR system testing?

3. Feature extraction and system testing

- DyViS corpus (Nolan et al. 2009): 94 young RP males recorded twice
- task 1: police interview, task 2: phone conversation (studio quality) with accomplice

MFCC and LTFD extraction

- samples divided into vowels (Vs) and consonants (Cs) using StkCV
- samples reduced to 60s of Vs per speaker
- 20ms frames/10ms shift (50% overlap) = 6000 frames per speaker/sample
  - **12 MFCCs, 12 Δs, 12 ΔΔs**
  - **F1-F4 frequencies, F1-F4 Δs, F1-F4 bandwidths**
  - (M)LTFDs: Mel weighted LTFDs

VQ extraction

- modified version of Laver's Vocal Profile Analysis (VPA; San Segundo et al in press)
  - 25 supralaryngeal settings & 7 laryngeal settings
  - **Task 1:** subset of speakers based on errors made by the best (S-)ASR system
  - **Task 2:** agreed VPAs for 100 speakers (based on three raters' evaluations)

Likelihood ratio (LR)-based testing

- 94 speakers divided into training (31), test (31) and reference (32) sets
- same- (SS) and different-speaker (DS) LR-like scores computed for training and test sets (GMM-UBM, Reynolds et al 2001); calibration and fusion (Morrison 2013)
- systems evaluated via equal error rate (EER) and log LR cost function (C_{llr})

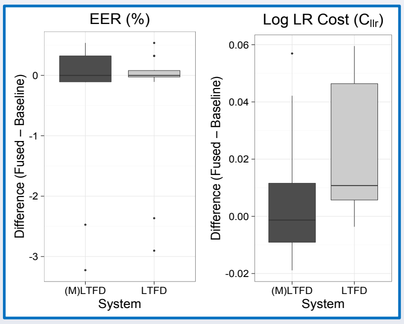
4. Results

Individual systems

- best individual system = **MFCCs+Δs+ΔΔs (EER = 3.23%, C_{llr} = 0.146)**
- best formant system = LTFDs+Bandwidths (EER = 6.45%, C_{llr} = 0.255)
  - Mel weighted LTFDs produced poorer performance (EER > 8%, C_{llr} > 0.3)

Fused systems

- 24 pairwise combinations of MFCC and (M-)LTFD systems tested



> 0 = fused system better than baseline

0 = no improvement

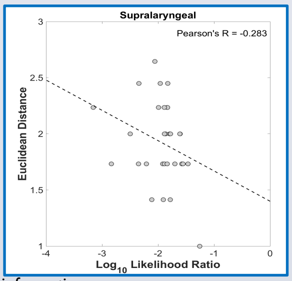
< 0 = fused system worse than baseline

- best system overall = **MFCCs+Δs+ΔΔs** and **LTFDs (EER = 3.23%, C_{llr} = 0.137)**

Evaluation of *errors* using supralaryngeal VQ

- best system produced 14 errors: 13 false acceptances (DS pair producing SS evidence) and 1 false rejection (SS pair producing DS evidence)
- 9/13 false acceptances involved speakers #67 and #72
  - is there anything about their supralaryngeal VQ profiles which might explain this?
  - non-neutral for **advanced tongue tip, fronted tongue body,** and **nasality**
  - settings shared by over 60% of the DyViS sample: so common as to be considered accent features for this group
  - easily confused with other speakers? *lambs* in the biometric menagerie?

- y-axis = Euclidean distance calculated between each test speaker's supralaryngeal VQ profile and the average (mode) profile for all 100 speakers
- x-axis = mean of the DS LLRs for each test speaker (i.e. every DS comparison they were involved in)



The role of laryngeal VQ

- misclassifications easily resolved using laryngeal VQ information
  - 8/13 false acceptances: differences of 2 or 3 scalar degrees (often neutral vs. non-neutral distinction) for at least 1 laryngeal setting
- misclassified pairs analysed blind by two forensic experts
  - correctly separated all 14 pairs
  - laryngeal VQ is a key feature

5. Discussion

- LTFDs consistently outperformed (M-)LTFDs
  - lower resolution representation of higher frequencies which are known to encode considerable speaker-specific information (e.g. F3)
- limited improvement in MFCC baseline when fused with formant information
  - MFCCs capture the same (and more) speaker-specific information as the formants
- supralaryngeal VQ captures some of the same information as MFCCs/LTFDs: unremarkable VQ speakers more likely to produce weak DS LLRs or false acceptances
- laryngeal VQ appears to capture orthogonal speaker-specific information – despite being problematic for the (S-)ASR, they are easily separated using auditory analysis

6. Conclusions

- understanding the relationships between different measures associated with different methods of analysis in FVC helps us to identify problematic cases and to better explain what information our systems capture (to lawyers and jurors)
- more work needed at the interface of different methods to further improve the validity and reliability of FVC evidence presented to the Courts
- research needed on forensically realistic recordings

BAAP 2018, University of Kent