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

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

1. Introduction

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

Three common methods of analysis

- **linguistic-phonetic**
- **automatic (ASR)**
- **semi-automatic (SASR)**

1.1. Increasing focus on integrating methods in research (IHASC: 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 Stade and Ors (2015))
- improvement in FVC system performance using combined methods

1.2. 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 vowels 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?
3. can supralaryngeal VQ explain the errors made by the (S-)ASR system?
4. what is the potential value of laryngeal VQ to (S-)ASR system testing?

3. Feature extraction and system testing

**MFCC and LTFD extraction**

- samples divided into vowels (V) and consonants (C) using 56CV
- samples reduced to 60s of 16 per speaker
- 20ms frames/10ms delta (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/VPA for 100 speakers (based on three raters’ evaluations)

**Likelihood ratio (LR)-based testing**

- 94 speakers divided into training (31), test (31) and reference (31) sets
- same (S)- and different-speaker (DS) LR-like scores computed for training and test sets
- systems evaluated via equal error rate (EER) and logLR cost function (Cll)

4. Results

**Individual systems**

- best in individual system = MFCC+Δs+ΔΔs (EER = 3.23%, Cll = 0.146)
- best formant system = LTFDs bandwidths (EER = 6.45%, Cll = 0.255)
  - Mel weighted LTFDs produced poorer performance (EER > 8%, Cll > 0.3)

**Fused systems**

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

- best system overall = MFCC+Δs+ΔΔs and LTFDs (EER = 3.23%, Cll = 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 h67 and h72
  - 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 DyVo sample so common as to be considered access features for this group
  - easily confused with other speakers’ limbs in the biometric menagerie?

- y-axis = Euclidian 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 LRs 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
  - 9/13 false acceptances differenced 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 LRs 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