Mapping across feature spaces in forensic voice comparison

the contribution of auditory-based voice quality to (semi-)automatic system testing

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}@york.ac.uk

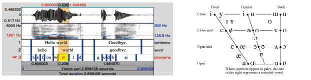
Arts & Humanities Research Council
voice and identity

Introduction

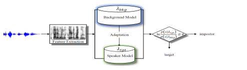
Forensic voice comparison (FVC) = offender (unknown) vs. suspect (known)

Three common methods of analysis

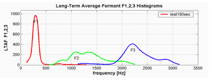
linguistic-phonetic


Hello world.
hello
world
Goodbye.
goodbye
sentence
word
phoneme
Visible part 2.898458 seconds
Total duration 2.898458 seconds
Front
Central
Back
Close
Close-mid
Open-mid
Open
Where symbols appear in pairs, the one to the right represents a rounded vowel.

automatic (ASR)



semi-automatic (SASR)



Integrating methods

- Increasing focus on the combination of methods
- Research: (H)ASR in NIST 2010 (Greenberg et al 2010), Gonzalez-Rodriguez et al (2014), Zhang et al (2013), Hughes, Foulkes and Wood (2016)
- Casework: Sweden and Germany (BKA)

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 combinations of methods

Features for analysis in this study

Voice quality (VQ): quasi-permanent vocal settings separated into 'supralaryngeal' and 'laryngeal' settings. Analysed by experts regularly in casework and considered one of the most useful linguistic-phonetic features for speaker separation (Gold and French 2011)

Mel frequency cepstral coefficients (MFCCs): rich representation of the Mel-weighted power spectrum, decoupling supralaryngeal and laryngeal information

Long term formant distributions (LTFDs): modelling formant values extracted automatically from all vocalic elements in the speech stream. Requires information about vowel boundaries, but not *segmental* in that all vowels are modelled together

Why these?

- Most commonly used features in each FVC method
- Shown to encode considerable speaker-specific information
- All, in principle, capture information about the supralaryngeal vocal tract

Research questions

1. How does the performance of MFCCs and LTFDs compare?
2. Does combination (fusion) of MFCC and LTFD systems improve performance over MFCCs 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?

Feature extraction and system testing

- DyViS database (Nolan et al. 2009): 100 young RP males recorded twice
- Task 1: police interview, Task 2: telephone conversation with accomplice

MFCC and LTFD extraction:

- Samples divided into vowels (Vs) and consonants (Cs) using StkCV
- Samples reduced to 60s of Vs per speaker (6 speakers removed)
- 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; see San Segundo et al submitted) scheme used = 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 sets of training (31), test (31) and reference (32) speakers
- Same- (SS) and different-speaker (DS) LR-like scores computed for training and test sets using GMM-UBM approach (Reynolds et al 2001)
- Score-level calibration and fusion using logistic regression (see Morrison 2013)
- Systems evaluated using equal error rate (EER) and the log LR cost function (C_{llr})

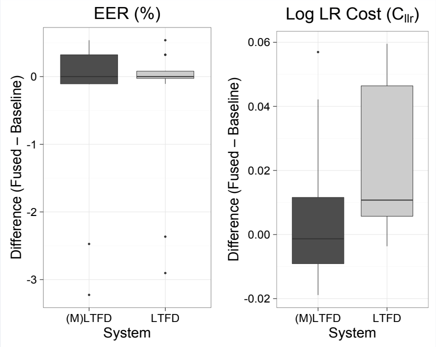
Results

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 weighting 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

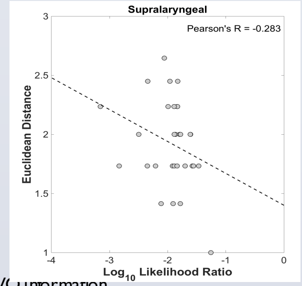
0 = no improvement

<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 (contrary-to-fact LLRs): 13 false acceptances (DS pair producing SS evidence) and 1 false rejection (SS pair producing DS evidence)
- 9 of the 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 by two forensic experts who produced LR-like scores
    - Able to correctly separate all 14 pairs
    - Laryngeal VQ described as a key feature

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 capturing the same speaker-specific information as the formants (in fact, the same and more, based on individual performance)
- Supralaryngeal VQ capturing some of the same information as the MFCCs/LTFDs: weak correlation suggesting 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

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

INTERSPEECH 2017
Situated Interaction
August 20-24, 2017 | Stockholm, Sweden
www.interspeech2017.org