The individual and the system: assessing the stability of the output of a semi-automatic forensic voice comparison system

University of York

J P French Associates
Forensic speech and acoustics laboratory

Vincent Hughes, Philip Harrison, Paul Foulkes, Peter French

Arts & Humanities Research Council

voice and identity

Introduction

- Semi-automatic (SASR) forensic voice comparison (FVC):
  - manual feature extraction: usually formants (LTFDs)
  - automatic modelling, scoring, evaluation
- Benefit = features are easier to explain to courts than ASR
  - clearer relationship between articulation and acoustics
- Previous work has shown good performance (e.g. [1,2,3]), but…
  - generally focus on matched conditions
  - analysis based on overall system performance

Research questions

1. How is the performance of a formant-based SASR system affected by **mismatched conditions**?
2. To what extent is performance affected by **degradation in transmission quality**?
3. How are **individual comparisons** affected? Can we predict which speakers will be more/less sensitive to mismatch and degraded quality?

Methods

Corpus

- 97 DyViS [4] speakers: suspect = Task 1, offender = Task 2
- Four versions of the offender sample:
  - high quality (HQ): original near-end sample
  - landline telephone (TEL): original far-end sample
  - high bit-rate mobile telephone (MOB_{HQ}) } [5]
  - low bit-rate mobile telephone (MOB_{LQ}) } [5]

Formant extraction

- 60 second samples of vowel-only material created
- 9 feature vector extracted from 20ms frames with 10ms shift: F1, F2 and F3 frequencies, deltas (Δs) and bandwidths

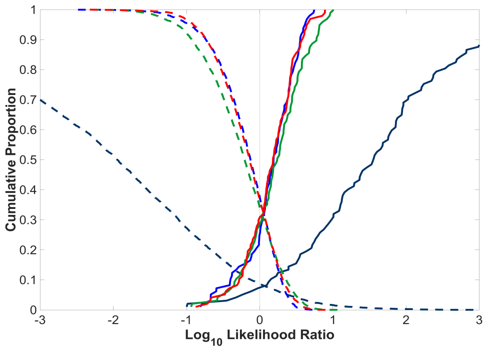
System testing

- Cross-validated same- (SS) and different-speaker (DS) scores computed using GMM-UBM [6] (using 8 Gaussians)
- Score-level logistic regression calibration [7] using cross-validation
- System validity: log LR cost (C_{llr}) and equal error rate (EER)
- Individuals analysed using means and standard deviations (SDs) of SS and DS LLRs: visualised using zooplots (see [8])

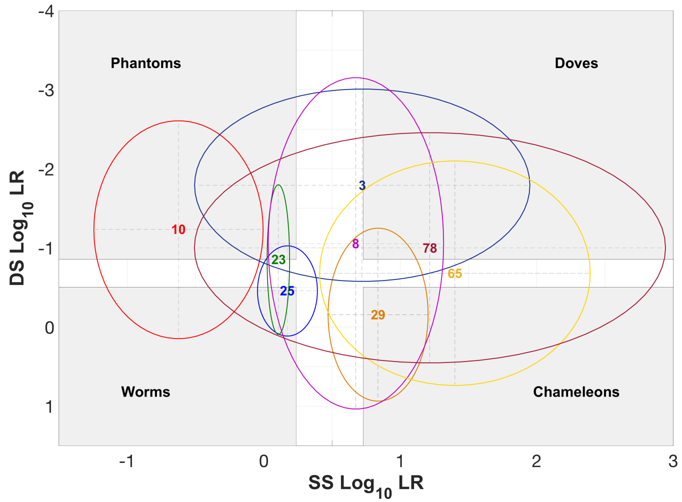
Results

System performance

	Suspect	Offender	EER (%)	C_{llr}
(1)	HQ	HQ	10.33	0.37
(2)	HQ	TEL	25.95	0.73
(3)	HQ	MOB_{HQ}	31.71	0.81
(4)	HQ	MOB_{LQ}	31.99	0.83



Individuals



Discussion

- Matched HQ condition provides the best overall performance
  - compare with 6.45% (EER) and 0.255 (C_{llr}) with the same recordings in [1] using F1, F2, F3, and F4
  - ∴ F4 provides useful information (where available)
- Decrease in performance as quality degraded
  - HQ < TEL < MOB_{HQ} < MOB_{LQ}
  - effect of bit-rate = relatively small

Predicting individual performance

- Some comparisons affected more/less by mismatch and degradation in quality (SDs as large as two orders of magnitude)
- Linear mixed effects models fitted to predict speakers' positions in the zoo space
  - **high SS mean == high DS mean**
  - **high means == high SDs**
  - **high mean F3 == high SS SD**

*voice quality and other formants were **not** significant

- F3 effect possibly due to default settings used
  - four formants tracked (LPC order = 12)
  - may have caused F3 measurement errors for speakers with high F3 (where F4 is outside the upper threshold for telephone transmission)

Conclusions

- Transmission mismatch between suspect and offender can have a substantial effect on SASR performance
- Considerable effect on LRs for individuals in terms of strength of evidence and variability
- Difficult to predict which comparisons will be most affected
  - but it may be possible to reduce effects by using channel- and speaker-specific (and possibly vowel-specific) formant settings

[1] Gold, E., French, P. & Harrison, P. (2014) Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proc. Meetings on Acoustics* 19. **[2]** Becker, T., Jessen, M. & Grigoras, C. (2008) Forensic speaker verification using formant features and Gaussian mixture models. *Proc. Interspeech*. pp. 1505-1508. **[3]** Jessen, M., Alexander, A. & Forth, O. (2014) Forensic voice comparisons in German with phonetic and automatic features using Vocalise software *Proc. AES*. pp. 28-35. **[4]** Nolan, F., McDougall, K., de Jong, G. & Hudson, T. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *IJSLL* 16: 31-57. **[5]** Alzqhoul, E., Nair, B. & Guillemin, B. (2015) An alternative approach for investigating the impact of mobile phone technology on speech. *Proc. World Congress on Engineering and Computer Science* vol. 1. **[6]** Reynolds, D., Quatieri, T. & Dunn, R. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10: 19-41. **[7]** Brümmer, N. et al. (2007) Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Trans. Audio Speech and Lang.* 15: 2072-2084. **[8]** Alexander, A., Forth, O., Nash, J. & Yager, N. (2014) Speaker recognition with tall and fat animals. *Paper at IAFPA 2014*. University of Zurich, Switzerland.