Modelling features for forensic speaker comparison

Vincent Hughes[1] (vh503@york.ac.uk), Erica Gold[1], Paul Foulkes[1], Peter French[1,2], Philip Harrison[1,2], Louisa Stevens[1,2], Colin Aitken[3], Tereza Neocleous[4], Víctor Medina[3] and Gary Napier[4]

[1]University of York, [2]J P French Associates, [3]University of Edinburgh, [4]University of Glasgow

Project aims

- Collaboration between forensic phoneticians and statisticians
- Funded by British Academy/Leverhulme Trust grant SG122381
- Aims to develop new statistical models that incorporate a broader array of phonetic variables into forensic speech comparison analyses and thus quantify forensic phonetic evidence more reliably.

Forensic speaker comparison

- Experts are presented with recordings of a criminal (e.g. threatening phone calls) and recordings of a suspect (e.g. police interview) and asked to assess the possibility that the recordings contain the voice of the same person.
- Where several features are analysed (e.g. vowels, consonants, fundamental frequency, ...) results for each individual variable must be combined into an overall measure of the strength of evidence.
- Likelihood ratios (LRs) have been recently applied to such forensic speaker comparison (FSC) problems. However, current applications generally fail to account for the complexity and inter-relatedness of variables.
- Voice evidence in the form of an LR therefore tends either to focus on a small subset of continuous acoustic variables (potentially overlooking other discriminatory variables), or to ignore the inter-relatedness of the variables and thus present a potentially misleading overall LR.

Our approach

- Develop statistical models based on all the variables measured.
- Consider both univariate (modelling one variable at a time) and multivariate models (modelling the inter-relatedness of the variables).
- Consider parametric distributions such as the normal and gamma, as well as nonparametric density estimation.
- Conduct simulation studies to validate the effectiveness of a particular variable or combination of variables for FSC.
- This approach is applied to obtain the evidential value of hesitation data.

Hesitation data

- Data extracted from the DyViS database of young (18-25), male speakers of Standard Southern British English.
- Variables measured: formant frequencies (F1 $<$ F2 $<$ F3) and duration (D) for the vowel portions of the hesitation marker 'um' from 75 speakers with 20 tokens per speaker.
- Each token can be treated as the same word. Hence a random effects model is appropriate, with two levels of variability: between and within speaker.

Simulation study 1: assuming normality

- Univariate and multivariate normal random effects models considering all combinations of F1, F2, F3 and D, with parameters estimated from a training set of 25 speakers.
- Same-speaker (SS) and different-speaker (DS) comparisons for a test set of 25 speakers: For SS comparison, the 20 tokens from each speaker in the test set were either split equally at random into a control and recovered sample (10:10) or ten were selected at random and split equally (5:5); for DS comparison either all 20 tokens from each speaker were compared to all 20 tokens from each of the other 24 speakers in the test dataset (20:20) or the 10 randomly selected tokens from each speaker were compared to the 10 tokens from each of the other speakers (10:10).

	Error %			
	SS (5:5)	DS (10:10)	SS (10:10)	DS (20:20)
F1	8.0	38.3	8.0	34.0
F2	4.0	29.0	4.0	23.0
F3	8.0	31.7	0.0	27.0
D	8.0	42.0	4.0	39.3
F1, F2, F3	4.0	8.0	0.0	6.3
F2, F3, D	4.0	12.0	0.0	6.7
F1, F2, F3, D	4.0	7.3	0.0	4.7

Table 1: Simulation study 1 rates of misleading evidence.

Results

The lowest misleading evidence rates were obtained for the multivariate model which includes F1, F2, F3 and D (Table 1). The error rates tend to be lower when more variables are considered, and when there are more tokens available per person (10:10 vs 5:5 for SS and 20:20 vs 10:10 for DS).

Simulation study 2: other distributions

- Gamma and truncated normal distributions are considered for the non-negative difference F2−F1, which captures the frontness of the vowel (Figure 1).

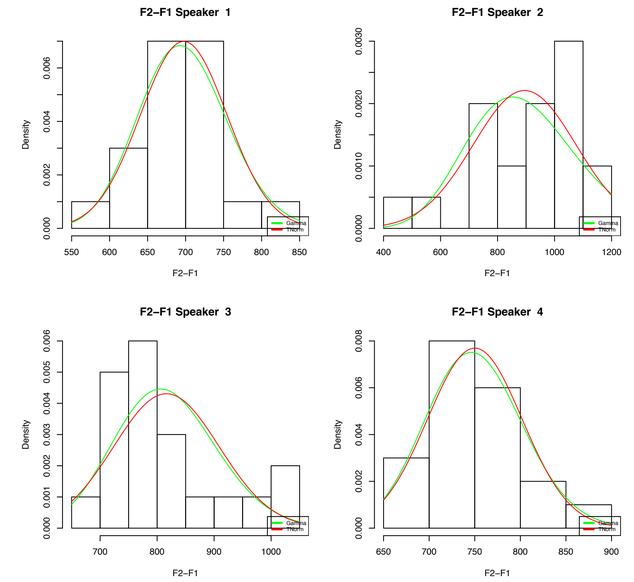

Figure 1: Histogram and density estimates for F2-F1 for speakers 1 to 4

- Alternatively, univariate and bivariate models for F1 and F2 using kernel density estimation (KDE) can be fitted.
- For SS comparisons, the 20 tokens are split into the first 10 and the second 10 tokens. For DS comparisons, all 20 tokens in each group are compared with all 20 tokens in the other 74 groups. The rates of misleading evidence are given in Table 1.

	Error %	
	SS (10:10)	DS (20:20)
F1, F2 (normal)	12.0	12.9
F1, F2 (KDE)	12.0	13.1
F2−F1 (gamma)	9.3	27.2

Table 2: Simulation study 2 rates of misleading evidence.

Ongoing work

Identify the most effective model and distributional assumptions for the hesitation data, considering post-processing (calibration) of LRs to improve performance, and produce recommendation on which variable(s) are the most effective.