Models for forensic speaker comparison

Tereza Neocleous[1] (tereza.neocleous@glasgow.ac.uk), Gary Napier[1] , Colin Aitken[2], Paul Foulkes[3], Peter French[3,4], Erica Gold[5], Philip Harrison[3,4], Vincent Hughes[3], Louisa Stevens[3,4]

[1]University of Glasgow, [2]University of Edinburgh, [3]University of York, [4]J P French Associates, [5] University of Huddersfield

Forensic voice 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 feature must be combined into an overall measure of the strength of evidence.
- Likelihood ratios have been recently applied to such forensic speaker comparison (FSC) problems.
- The likelihood ratio (LR) can be defined as

$$LR = \frac{p(\text{features}|\text{same speaker})}{p(\text{features}|\text{different speaker})}$$

where $p(\cdot)$ is a probability density or mass function to be estimated.

- Estimation of $p(\cdot)$ is one of the statistical issues in FSC.
- Voice evidence evaluation: LR values greater than 1 support the same speaker hypothesis, and LR values less than 1 support the different speaker hypothesis.

Statistical modelling aspects

- Several features are measured, but phonetic databases are typically small.
- How best to model the data?
- Assess model performance based on rates of misleading evidence.

Approach 1: Modelling one feature at a time

\+ Easy to implement, even in situations where there are many more features than observations (e.g. automatic speaker recognition systems).
\- Does not take into account correlations between features.
\+ Possible to combine the evidential value of multiple features by assuming independence (overall LR equals sum of LRs from each feature)...
\- ... but that inflates the value of evidence.
\+ Alternatively, post-processing techniques could be used to combine the evidential value of multiple features (e.g. logistic regression).
\- Results from such methods are less interpretable.

Approach 2: Multivariate modelling

Uses a multivariate distribution to model several features simultaneously.

\+ Takes into account correlations.
\- Not always possible to implement with small datasets.
\- Nonparametric estimation not stable for high-dimensional data.

Approach 3: Converting to scores

Uses a distance or similarity measure instead of individual features.

\+ Easy to implement.
\+ No need for post-processing (combining LRs) as a single LR is produced.
\- Information loss can occur when reducing the data from features to scores, which in turn can result in lower strength of evidence.

Application to hesitation data

- Data extracted from the DyViS database of young (18-25), male speakers of Standard Southern British English.
- Features 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 design

- Feature-based analysis (Approaches 1 & 2): Univariate and multivariate normal random effects models considering all combinations of F1, F2, F3 and D were fitted, 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 split equally; for DS comparison all 20 tokens from each speaker were compared to all 20 tokens from each of the other 24 speakers in the test dataset. The LR was computed assuming normal between- and within-speaker distribution.
- Score-based analysis (Approach 3): The Euclidean distance is calculated between all SS and DS pairs of observations. Density estimates of the score distributions are shown in Figure 1. The LR is computed as the ratio of the two densities.
- Model performance criterion: the percentage of time SS comparisons return LR < 1 and DS comparisons return LR > 1.

Results and discussion

Error rates for selected feature-based models are shown in Table 1.

	Error %	
	SS	DS
F1	22.5 (7.3)	37.0 (4.4)
F2	6.8 (4.0)	23.2 (3.4)
F3	13.5 (6.7)	27.3 (4.0)
D	14.2 (6.9)	35.1 (5.4)
F1, F2 (multiv)	4.0 (3.4)	26.3 (4.4)
F1, F2 (indep)	4.5 (3.7)	13.0 (2.5)
F1, F2, D (multiv)	0.4 (1.3)	19.0 (4.3)
F1, F2, D (indep)	3.8 (3.2)	7.9 (2.0)
F1, F2, F3, D (multiv)	0.8 (1.7)	14.0 (4.1)
F1, F2, F3, D (indep)	4.2 (3.4)	4.3 (1.8)

Table 1: Rates of misleading evidence for feature-based models. Each rate is an average of 100 replicates with the standard deviation shown in brackets.

- Univariate models perform poorly but combining information from multiple features improves error rates.
- Combining LRs assuming independence appears to perform better than multivariate normal models, possibly due to the small training dataset and low correlations between features.
- Using all four features performs best, although the model with just F1, F2 and duration also performs reasonably well. This is relevant because F3 is not always available in casework data.
- Score-based models perform poorly due to the overlap between same- and different-speaker distributions seen in Figure 1.

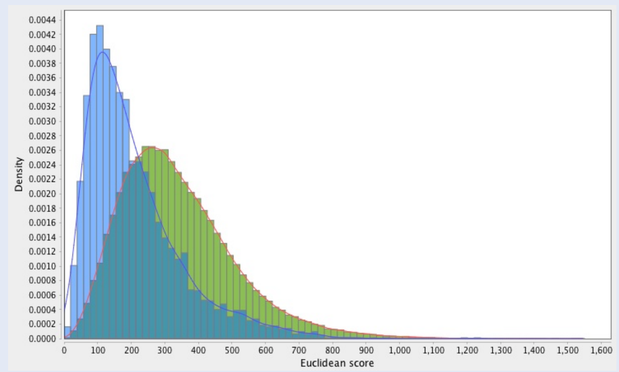

Figure 1: Score distribution for same (blue) and different (green) speaker pairs

This work was supported in part by British Academy/Leverhulme Trust grant SG122381.