Sociolinguistic considerations in the default definition of the ‘relevant population’ when computing numerical likelihood ratios

Vincent Hughes[1,2] *and Paul Foulkes*[1,2,3]
[1]*Department of Language and Linguistic Science, University of York, UK.*
[2]*New Zealand Institute of Language, Brain and Behaviour,*
University of Canterbury, New Zealand.
[3]*J P French Associates, York, UK.*
{vh503|paul.foulkes}@york.ac.uk

To estimate the strength of forensic voice comparison (FVC) evidence using a likelihood ratio (LR) it is essential to assess not only the similarity across the known and disputed samples, but also their typicality in the *relevant population*. Previous attempts to define a default *relevant population* for FVC have generally ignored many of the complex dimensions of within- and between-speaker variation widely documented in sociophonetics (e.g. Foulkes and Docherty 2006). Instead, FVC studies have typically gone no further than controlling for (assumed) native language and speaker sex (e.g. Rose 2004; Morrison et al 2012). However, little is known about the extent to which LRs are affected by other sociolinguistic factors that may be ‘logically relevant’ (Kaye 2004) and which are known to affect the frequency of parameters in the population. This paper investigates the effects of different definitions of the *relevant population* on absolute strength of evidence and system performance. Our focus in this study is variation in the socio-economic background and age of the speakers included in the analysis.

Dynamic time-normalised formant data (McDougall 2004) for the FACE vowel (/eɪ/, in *face, say, eight*, etc.) were auto generated from a subset of the ONZE (Origins of New Zealand English) corpus (Gordon et al. 2007). FACE was chosen for it’s expected levels of between-speaker variation in NZE, with Hay et al. (2008:43) claiming that “as well as changing over time, different pronunciations … have strong social connotations”. Procedures were implemented to correct formant tracking errors and a sub-section of the data was inspected manually. The first three formants of each token were fitted with quadratic polynomials and the coefficients used as input data when computing LRs. The data were initially divided by sex, and then into two experimental conditions according to:

(i) Socio-economic class (professional/ non-professional – following the coding used in ONZE)

(ii) Age (older/ younger – born before or after 1960, based on the bimodal split in the data)

For each condition, independent homogeneous (i.e. all professional/ all young) development and test sets containing 22 speakers were created. Cross-validated same-speaker (SS) and different-speaker (DS) LR scores were computed Aitken and Lucy’s (2004) MVKD formula. Typicality was assessed using three 40-speaker reference sets per experimental condition: (a) ‘tailored’ – direct class/age match with test set, (b) ‘mismatch’ – all non-professional/old, (c) ‘mismatch’ – containing equal age/class mix. Logistic-regression calibration weights were determined from scores for the development data and then applied to scores for the test data. Accuracy within conditions was assessed using the log likelihood ratio cost function (C_{llr}) (Brümmer and du Preez 2006). Precision of LRs between conditions (according to sex) was assessed using the non-parametric 95% credible interval (95% CI) (Morrison 2011).

Across all conditions absolute strength of evidence, system accuracy and system precision were affected to some extent by variability in the reference data. For both males and females, the mean, variance and range of LRs were overestimated in the 'mismatch' conditions compared with the 'tailored' baseline (with the exception of SS pairs in the class-female condition). In all cases, LRs computed using a 'mixed' reference set were more similar to those using the 'tailored' set, compared with those using the 'mixed' set. In terms of accuracy, for both class-male and age-male C_{llr} was highest using the 'mixed' set, although the absolute difference in performance between the sets was relatively small. For females the differences in C_{llr} were more marked with the 'mismatch' set overestimating accuracy by around 0.2 compared with the 'tailored' baseline. In all cases the 'tailored' set achieved the highest C_{llr} (class: 0.75/ age: 0.773). For females, mean 95% CI in the class condition was higher across the three reference sets (1.327) than for males (0.972), whilst the opposite was true in the age condition (females: 1.371/ males: 1.14).

The results of this study highlight a number of important issues for the quantification of strength of evidence using the numerical LR. First, sources of between-speaker variation affect LR output in different ways and to different extents, and it is essential to understand and acknowledge the sociophonetics of the parameters under investigation when conducting casework. Secondly, across both conditions the wrong definition of a 'logically relevant' factor (i.e. mismatch) had a more detrimental effect on LR output and system performance compared with no control (i.e. mixed). Finally, the differences between the male and female groups reveal the complex interaction between 'logically relevant' factors which may need to be controlled in a FVC case.

References

Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 54, 109-122.

Brümmer, N. and du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language* 20(2-3): 230-275.

Foulkes, P. & Docherty, G.J. (2006) The social life of phonetics and phonology. *Journal of Phonetics* 34: 409-438.

Gordon, E., Maclagan, M. and Hay, J. (2007) The ONZE corpus. In Beal, J. C., Corrigan, K. P. and Moisl, H. (eds.) *Models and Methods in the Handling of Unconventional Digital Corpora: Volume 2, Diachronic* Corpora. London: Palgrave. 82-104.

Hay, J., Maclagan, M. and Gordon, E. (2008) New Zealand English. Edinburgh: Edinburgh University Press.

Kaye, D. H. (2004) Logical relevance: problems with the reference population and DNA mixtures in *People v. Pizarro*. *Law, Probability and Risk* 3: 211-220.

McDougall, K. (2004) Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law* 11(1): 103-130.

Morrison, G. S. (2011) Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice* 51(3): 91-98.

Morrison, G. S., Ochoa, F. and Thiruvaran, T. (2012) Database selection for forensic voice comparison. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore*. International Speech Communication Association.

Rose, P. (2004) Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective. Keynote paper, *Forensic Speaker Recognition Workshop, Speaker Odyssey '04*. 31 May - 3 June 2004, Toledo, Spain. 3-10.