Effects of variation on the computation of numerical likelihood ratios for voice comparison

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

The likelihood ratio (LR) is rapidly becoming established as the "logically and legally correct" framework for the assessment and presentation of expert forensic evidence (Rose and Morrison 2009: 143). Applied to forensic voice comparison (FVC) the competing propositions of the prosecution and defence are represented as an assessment of the similarity and typicality of features across a pair of suspect and offender samples. The typicality element is dependent on patterns in the *relevant population* (Aitken and Taroni 2004: 206) and is quantified relative to a sub-section of that population when computing numerical LRs. The appropriate definition and delimitation of the relevant population is therefore essential if experts are to provide a reliable estimation of the strength of a given strand of FVC evidence.

In forensic casework, the analyst cannot know for certain the relevant population of which the offender is a member, since by definition the identity of the speaker is unknown. Pragmatic decisions must therefore be made regarding the definition of this population in order to compile a useable set of reference data. However, LR studies display a lack of consensus over the size and scope of the reference sample. Furthermore, they generally take account of only a small number of the linguistic, social and stylistic factors that may influence natural speech production: speaker sex, broadly defined dialect, and intra-speaker variation judged via non-contemporaneous but stylistically homogeneous samples. Factors such as these may affect LR calculations, and thus impact on the reliability of the overall analysis.

This study therefore assesses the effects on LRs when key dimensions are varied: (i) the number of reference speakers, (ii) the number of tokens per speaker in the reference distribution, and (iii) the dialect match between the suspect-offender material and the reference data. Using polynomial fittings of F1 and F2 trajectories from spontaneous GOOSE (/uː/) vowels, LR comparisons were performed against a New Zealand English (NZE) reference set of up to 120 speakers and up to 13 tokens per speaker. GOOSE is expected to display some limited geographically defined variation, with F1 and F2 more closely linked with patterns in the speech community than higher formants (Clermont et al 2008). Given that variable input necessarily affects the numerical output, the magnitude and systematicity of LR differences are assessed with reference to Champod and Evett's (2000: 240) verbal scale.

Mean same-speaker (SS) LRs were found to be relatively robust to variability on dimension (i), displaying minor fluctuations within the category of 'limited support' for the prosecution. Consistent with Ishihara and Kinoshita (2008), only when the size of the reference data was small (fewer than 15 speakers) did the strength of evidence change categorically, at which point mean SS LRs incorrectly offered support for the defence. Different-speaker (DS) pairs were more sensitive to such variation with a categorical increase on the verbal scale of strength of evidence found with fewer than 45 speakers. Variance and severity of error (C_{llr}) were also continually reduced with the inclusion of more data. On dimension (ii), mean and variance of SS LRs were consistent, even when the number of tokens per speaker in the reference data was reduced to as few as two. However, mean DS LRs were found to be around 10^4 times greater when using two tokens per speaker compared with 13 tokens per speaker, thus overestimating the strength of evidence when fewer tokens were included in the analysis.

The dialect composition of the reference distribution was varied to test Rose's default assumption that, in the absence of a specific alternative proposition, the relevant population should be "same-sex speaker(s) of the language" (2004: 4). LR computation was performed on mock suspect and offender data from one matching (NZE) and three mismatching test sets (Manchester, Newcastle and York). In the absence of differences in within-speaker variation, SS LRs proved to be on average 3.3 times higher for the mismatch sets: an overestimation of the strength of evidence equivalent to the difference between 'limited' and 'moderate' support for the prosecution. Mismatch DS comparisons also displayed high levels (58-71%) of contrary-to-fact support for the prosecution.

However, when limiting the analysis to F2 trajectories only, no categorical difference in the verbal strength of evidence was found between the matching and mismatching test sets for SS pairs. In each case the mean strength of evidence was equivalent to 'limited' support for the prosecution. These results indicate that the removal of certain regionally defining acoustic information (F1) may reduce the effect of dialect divergence between the evidential and reference data. The positive practical implication of this is that reliable LR computation may be possible using a broader definition of the relevant population for features which are expected to carry less dialect-specific information.

References

Aitken, C. G. G. and Taroni, F. (2004) Statistics and the evaluation of evidence for forensic scientists (2nd edition). Chichester: John Wiley.

Champod, C. and Evett, I. W. (2000) Commentary on A.P.A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification'. *Forensic Linguistics,* **7(2)**, 238-243.

Clermont, F., French, J. P., Harrison, P. and Simpson, S. (2008) Population data for English spoken in England: a modest first step. Paper presented at the International Association for Forensic Phonetics and Acoustics (IAFPA) Annual Conference. 21-23 July 2008, Swiss Federal Institute of Technology, Lausanne.

Ishihara, S. and Kinoshita, Y. (2008) How many do we need? Exploration of the Population Size Effect on the performance of forensic speaker classification. Paper presented at the 9th Annual Conference of the International Speech Communication Association (Interspeech). Brisbane, Australia. 1941-1944.

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.

Rose, P. and Morrison, G. S. (2009) A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech, Language and the Law,* **16(1)**, 139-163.