Investigating the effects of sample size on numerical likelihood ratios using Monte Carlo simulations

Vincent Hughes[1,2]
[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.*
vh503@york.ac.uk

An essential element of the likelihood ratio (LR) as applied to forensic voice comparison (FVC) is an assessment of the within- and between-speaker typicality of values for the linguistic-phonetic parameters of interest within the *relevant population*. To estimate typicality it is necessary to generate a sample of the *relevant population*, commonly termed the reference data. A significant practical issue is how large such a sample needs to be in order for the quantification of strength of evidence to be reliable. A limited set of studies show consistent findings that LRs are generally unstable with small numbers of speakers (Ishihara and Kinoshita 2008) and tokens per speaker (Hughes and Foulkes 2012). However, such studies have not investigated the upper limits at which LR performance becomes robust to sample size.

Monte Carlo simulations (MCS) offer a means of investing this issue without requiring an extremely large database or the extraction of a considerable amount of raw phonetic information. MCS involve generating synthetic values from a distribution of raw data such that the properties of the resulting distribution are delimited by the original data. The present study uses MCS to investigate how the number of reference speakers and number of tokens per reference speaker affect LR output, where the most precise estimate of strength of evidence (i.e. the closest to the 'true' LR, Rose 2012) is that based on the largest amount of reference data. Local articulation rate (AR) data, collected as part of Gold (in preparation), was used as raw input. The data consist of the first 26 'tokens' (phonological syllables per memory stretch) from 99 male DyVis speakers. Local AR was chosen as it is univariate and does not require modeling of complex multidimensional correlations between features of the same parameter.

20 speakers were extracted at random to function as a test set. Data from the remaining 79 speakers were used to generate a reference data set of up to 1000 speakers with up to 100 tokens per speakers. In the first experiment, same- (SS) and different-speaker (DS) LRs were computed using Aitken and Lucy's (2004) MVKD formula, starting with 10 reference speakers and increasing one at a time up to 1000. In the second experiment, LRs were computed as a single token per speaker was added to the reference data, beginning with two tokens and ending with 100. At each stage in both experiments the equal error rate (EER) and log-LR cost function (C_{llr}) (Brümmer and du Preez 2006) were calculated.

Despite Rose's (2012) finding that mean SS LRs stabilised and were very close to a 'true' LR by around 30 speakers, the results for the first experiment reveal that the mean SS $\log_{10}$ LR became negative after 104 speakers, thus offering contrary-to-fact support for the defence. After this point the SS mean becomes relatively stable, displaying only minor fluctuations within the range of 'limited' support for H_d (Figure 1). Similarly, mean DS strength of evidence displayed a categorical increase from 'limited' to 'moderate' support for H_d with greater than 114 reference speakers. Only

with greater than 200 speakers did EER and C_{llr} begin to stabilise. In the second experiment both SS and DS LRs were found to be relatively robust to the number of tokens per reference speaker, displaying marked fluctuation in performance only with very small numbers of tokens (<5 per speaker).

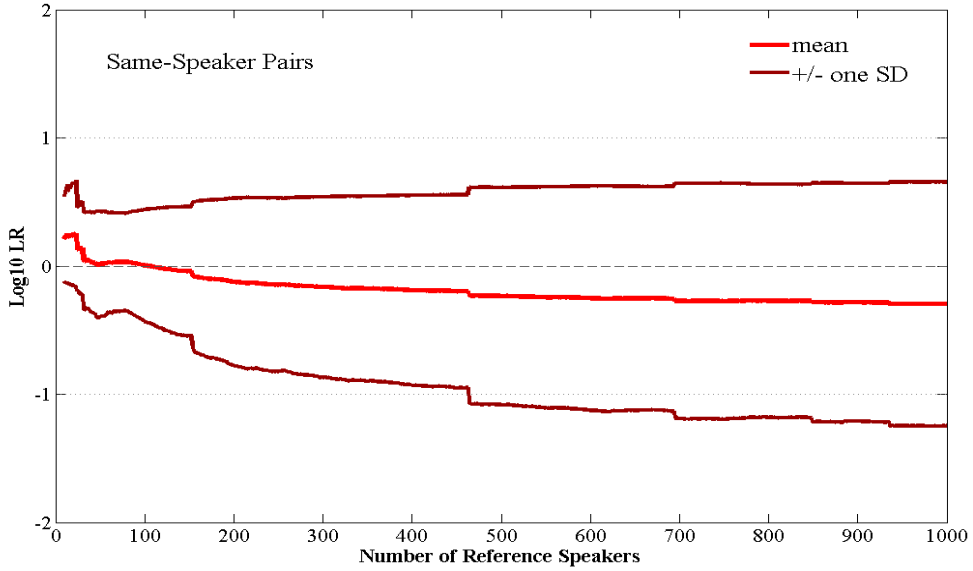

Figure 1 Mean and variance of same-speaker $\log_{10}$ LRs as a function of the number of reference speakers

The results have three important implications for the application of the numerical LR approach to FVC. First, different linguistic-phonetic parameters may behave in different ways with regard to the size of the sample. Secondly, such behaviour is, in part, a consequence of the speaker-discriminatory value of the parameter itself. Finally, the variability in system performance, even with large amounts of reference data, suggests that relying on the lowest C_{llr} or EER may in some situations be inappropriate in casework. Rather, it may be necessary to conduct some form of pre-testing to assess the relative stability of the LRs from the system.

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.

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.

Gold, E. (in preparation) Calculating likelihood ratios in forensic speaker comparison cases using phonetic and linguistic features. PhD Dissertation, University of York, UK.

Hughes, V. and Foulkes, P. (2012) Effect of variation on the computation of numerical likelihood ratios for forensic voice comparison. Paper presented at the International Association of Forensic Phonetics and Acoustics conference (IAFPA). Santander, Spain.

Rose, P. (2012) The likelihood ratio goes to Monte Carlo: the effect of reference sample size on likelihood-ratio estimates. Paper presented at UNSW Forensic Speech Science conference. Sydney, Australia.