Effects of the imbalance in the log LR cost function (*C*llr)

Vincent Hughes[1]
[1]*Department of Language and Linguistic Science, University of York, UK.*
vincent.hughes@york.ac.uk

Across the forensic sciences there have been increasing calls for the likelihood ratio (LR) to be used to evaluate the strength of expert comparison evidence and that the validity and reliability of systems used to compute a LR should be demonstrated empirically (NRC 2009; Morrison 2014). Within the field of forensic voice comparison (FVC) the log LR cost function (*C*llr; Brümmer and du Preez 2006) is commonly used as a measure of validity:

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}} log_2\left(1+\frac{1}{LR_{ss_i}}\right)+\frac{1}{N_{ds}}\sum_{i=1}^{N_{ds}} log_2\left(1+\frac{1}{LR_{ds_i}}\right)\right)$$

The *C*llr is a gradient measure which penalises the system based on the magnitude of *contrary-to-fact* LRs (SS LLRs < 0, DS LLRs > 0). *C*llr is philosophically consistent with the LR framework since it is not based on posterior accept-reject decisions, unlike the equal error rate (EER). *C*llr is commonly used to compare the performance of different systems using development and test sets containing same speaker (SS) and different speaker (DS) pairs evaluated against a set of speakers representative of the relevant population. Within LR-based testing there is an imbalance in the number of SS (= N speakers) and DS (minimally = (N^2-N)/2 speakers) comparisons performed. Since *C*llr is the summation of two cost functions derived from the SS and DS comparisons separately, this imbalance will be reflected in system validity. This paper presents two experiments which consider the effects on *C*llr of the inherent imbalance in the number of SS and DS comparisons in LR-based testing.

Firstly, sets of hypothetical, normally distributed SS (20) and DS (180) LLRs were simulated. Two systems were then compared: in System(1) all LLRs were increased by 1 and in System(2) all LLRs were decreased by 1. Therefore, both systems had the same EER (10%) but different proportions of false hits and misses.

Table 1. *C*llr values based on hypothetical systems with higher proportion of false hits (1) and misses (2) using imbalanced N SS and N DS comparisons.

	***C*llr**
System(1)	0.5010
Systems(2)	0.4499

Table 1 shows that System(1), with the higher proportion of false hits, was penalised to a greater extent (i.e. higher *C*llr) than System(2), with the higher proportion of misses. This indicates that, all else being equal, *C*llr inherently favours systems with higher proportions of misses than false hits. Secondly, sets of hypothetical LLRs with different EERs (6%, 10%, 16%, 22%) were generated and *C*llr values calculated shifting the EER thresholds between -3 and +3 using imbalanced (20 SS/180 DS) and balanced (180 SS/180 DS) sets. Maximally the imbalanced systems generated *C*llr values of 0.63 less than the balanced systems. However, the extent of these effects is dependent on the LLR value at the threshold for EER and the EER itself.

This paper also explores the relevance of validity measures (both EER and *C*llr) in the context of real casework. In terms of interpreting the LR presented to the court, the LR of the LR or the credible interval (CI) are considered much more useful to the trier-of-fact.

References

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

Morrison, G. S. (2014) Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. *Science and Justice*, **54(3)**, 245-256.

National Research Council (2009) Strengthening forensic sience in the United States: a path forward. http://www.nap.edu/catalog.php?record_id=12589. Accessed: 27th May 2014.