Calculating the reliability of likelihood ratios: Addressing modelling problems related to small *n* and tails

Geoffrey Stewart Morrison[1*], *Felipe Ochoa*[1], *Jonas Lindh*[2]

[1]Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia

[2]Division of Speech and Language Pathology, Department of Clinical Neuroscience and Rehabilitation, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Sweden

*geoff-morrison@forensic-voice-comparison.net

In forensic speech science we are often faced with the problem of having a relatively small amount of data which is also multivariate and distributionally complex. This results in a serious problem exactly in the scenario where potentially large strengths of evidence could be obtained, i.e., when the trace data are on a tail of the distribution which models either the prosecution or defence hypothesis and a large magnitude log likelihood ratio is calculated. By definition the sampling of a distribution is sparse on its tails and this problem is compounded if the model is trained on a small amount of data – small fluctuations in the training data can lead to large changes in the calculated likelihoods on the tails and thus large changes in the calculated likelihood ratios for trace data on the tails. Large-magnitude calculated log likelihood ratios are therefore inherently unreliable.

We illustrate this problem and explore a way of dealing with it using an extreme example based on data from a disputed-utterance case. In this Swedish case a word on an audio recording was disputed as being either the name "Tim" [tʰɪm] or the pronoun "dom" [dɒm] (they). The recoding also contained 16 undisputed tokens of "Tim" and 29 of "dom" spoken by the same speaker. VOT, F1, and F2 were measured for each undisputed word and for the disputed word. The [VOT, F1, F2] vector of the disputed word was near the middle of the "dom" model distribution but far out on a tail of the "Tim" model distribution (see Fig. 1), and the calculated likelihood ratio was 10^{77}! This is a ridiculously large number (the number of stars in the observable universe is around 10^{22}) and cannot possible be supported by the small amount of data used to build the models.

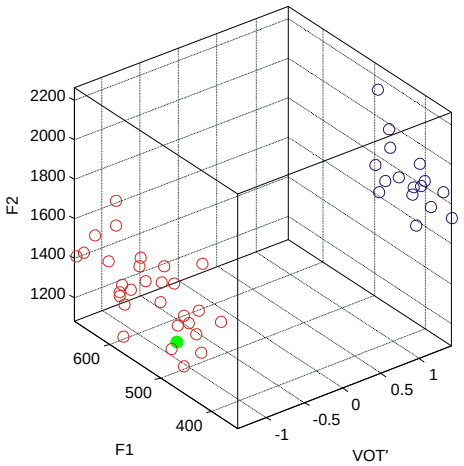

Figure 1: "Tim" data (blue circles), "dom" data (red circles), and disputed-word data (green dot).

Monte Carlo simulations were conducted using the mean vectors and covariance matrices from the case data as population parameters for generating simulated data. Ten thousand sets of 16 simulated "Tim" and 29 simulated [VOT, F1, F2] vectors were generated. For each set, models were built and a likelihood ratio calculated for the [VOT, F1, F2] vector of the disputed word. In order-of-magnitude bins, the modal value for the Monte Carlo likelihood ratios was 10^9 and the 1st and 5th percentiles were 10^6 and 10^9 (Fig. 2).

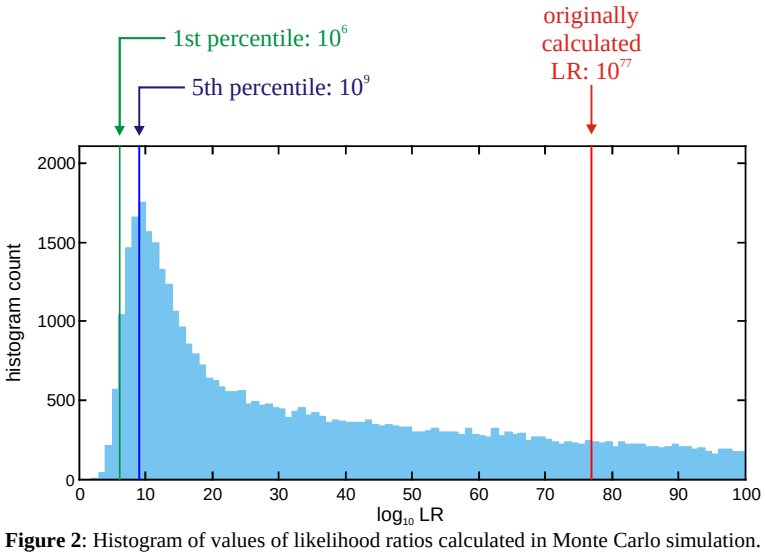

Figure 2: Histogram of values of likelihood ratios calculated in Monte Carlo simulation.

Ridiculously large calculated likelihood ratios which cannot possibly be supported by the small amounts of data available should not be reported. Instead we propose that making a statement such as "I'm 99% certain that the likelihood ratio is at least 10^6" would be appropriate.

Although we have used an extreme example, modelling reliability is in fact a general issue which can potentially be a problem for any model in any case, and it may be appropriate to assess this sort of modelling reliability in every case as a matter of course.