Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Estimating the Precision of the Likelihood-Ratio Output of a Forensic-Voice-Comparison System

Geoffrey Stewart Morrison[1,2], Tharmarajah Thiruvaran[2], and Julien Epps[2,3]

[1]School of Language Studies, Australian National University, Canberra, ACT 0200, Australia
[2]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia
[3]National ICT Australia (NICTA), Australian Technology Park, Sydney, NSW 1430, Australia
geoff.morrison@anu.edu.au, thiruvaran@student.unsw.edu.au, j.epps@unsw.edu.au

Abstract

The issues of validity and reliability are important in forensic science. Within the likelihood-ratio framework for the evaluation of forensic evidence, the log-likelihood-ratio cost (C_{llr}) has been applied as an appropriate metric for evaluating the accuracy of the output of a forensic-voice-comparison system, but there has been little research on developing a quantitative metric of precision. The present paper describes two procedures for estimating the precision of the output of a forensic-comparison system, a non-parametric estimate and a parametric estimate of its 95% credible interval. The procedures are applied to estimate the precision of a basic automatic forensic-voice-comparison system presented with different amounts of questioned-speaker data. The importance of considering precision is discussed.

1. Introduction

1.1. Concern about accuracy and precision in forensic science

Recently there has been a great deal of concern in forensic science about validity and reliability [1–4]. The National Research Council report to Congress on Strengthening Forensic Science in the United States [3] urged that procedures be adopted which include "the reporting of a measurement with an interval that has a high probability of containing the true value; . . . [and] the conducting of validation studies of the performance of a forensic procedure" (p. 121); the latter requiring the use of "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23).

1.2. Accuracy

In statistics and scientific literature *validity* is synonymous with *accuracy* and *reliability* with *precision*; however, in judicial and forensic-science literature reliability has often been discussed without explicit definition, or has been defined in terms of a measure of validity: *classification-error rates*, i.e., the proportion of same-origin comparisons in a test set which are classified as different-origin (misses), and the proportion of different-origin comparisons which are classified as same-origin (false alarms).

If one accepts that the *likelihood-ratio framework* is the correct framework for the evaluation of forensic comparison evidence [5–18], then a metric such as classification-error rate, based on a hard-thresholding of posterior probabilities, is not an appropriate measure of accuracy (by extension, this is also true for equal error rate, EER). Rather, an appropriate metric should be based on likelihood ratios (LRs) and should be continuous in nature – an LR which provides greater support for a contrary-to-fact hypothesis should attract a heavier penalty than one which provides more limited support for the contrary-to-fact hypothesis, since the former has a greater potential to contribute to a miscarriage of justice. An appropriate measure of accuracy, developed for use in automatic speaker recognition [19, 20] and subsequently applied in forensic voice comparison, e.g., [14, 21], is the *log-likelihood-ratio cost* (C_{llr}), which, at least in automatic speaker recognition, may now be considered a standard metric of the accuracy of a system which outputs LRs.

1.3. Precision

In addition to accuracy, however, it is also important to consider precision [22, 23]. Imagine two systems that are assessed as having the same accuracy and when tested on a particular pair of objects multiple times give the same average $\log_{10}$(LR) of −2, but the test results on one system have a wide range of LR output values leading to an estimated 95% credible interval, in $\log_{10}$(LR), of ±0.1 whereas the other has an estimated 95% credible interval of ±3. The former system (with a 95% LR credible interval for this pair of objects ranging from 79 to 126 in favor of the different-origin hypothesis) would be preferred over the latter (with a 95% LR credible interval for this pair of objects ranging from 100 000 in favor of the different-origin hypothesis to 10 in favor of the same-origin hypothesis). The former, more precise, system would be much more useful in assisting the trier of fact to weigh the forensic-comparison evidence as part of making their ultimate decision as to the guilt or innocence of the accused (the *trier of fact* is the judge, the panel of judges, or the jury, depending on the legal system).

The present paper describes and provides examples of the use of two procedures for calculating a metric of the precision of the LR output of a forensic comparison system. The metric is an estimate of the *95% credible interval* (CI) [24]. One procedure is non-parametric and the other parametric, and the examples are of their application to an automatic forensic-voice-comparison system. The aim of developing this metric is to allow forensic scientists to compare developmental systems, and to allow them to report the precision, as well as the accuracy, of the final system so that a judge can consider whether testimony based on the system should be admitted in court [1]. Finally, as part of their testimony, it would allow a forensic scientist to make a statement such as the following:

> Based on my evaluation of the evidence, I have calculated that one would be *X* times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been