Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework

Geoffrey Stewart Morrison [a,b]

[a] *School of Language Studies, Australian National University, Canberra, Australia*

[b] *School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia*

Abstract

In their recent introduction to forensic linguistics, Coulthard & Johnson (2007) include a portrayal of the likelihood-ratio framework for the evaluation of forensic comparison evidence (pp. 203–207). This portrayal includes a number of inaccuracies. The present letter attempts to correct these inaccuracies.

Keywords: likelihood-ratio framework; forensic linguistic comparison; forensic voice comparison

Introduction

In their recent introduction to forensic linguistics, Coulthard & Johnson [1] include a portrayal of the likelihood-ratio framework for the evaluation of forensic comparison evidence (pp. 203–207). Despite its brevity, 3.5 pages in a volume of 256 pages, it contains a number of inaccuracies (I do not mean to imply that these were deliberate on the part of the authors). In this letter I will attempt to correct these inaccuracies, lest the reader be left with an incorrect understanding of the likelihood-ratio framework. I will not provide a systematic description of the likelihood-ratio framework here, such descriptions can be found in several works including Aitken & Taroni [2], Balding [3], Champod & Meuwley [4], González-Rodríguez *et al.* [5], Lucy [6], and Rose [7,8].

"Defence likelihood ratio"

Coulthard & Johnson refer to a "defence likelihood ratio" (p. 205), but there is no such thing as a "defence likelihood ratio" (or a prosecution likelihood ratio). A likelihood ratio is the probability of observing the same evidence under two competing hypotheses, e.g., the probability of observing certain measurable acoustic differences between audio recordings of voices of questioned and known origin under the prosecution's claim that the speaker of the questioned voice is the same as the speaker of the known voice, the defendant, versus under the defence's claim that the questioned voice was produced by some speaker other than the defendant. Note also that there may be cases in which the defence claims same origin and the prosecution claims different origin, e.g., this may be the norm in cases of disputed authenticity of written texts. A general schema for a likelihood ratio in forensic applications is given in Equation 1.

$$\text{likelihood ratio} = \frac{p(\text{ observed difference between samples} \mid \text{same origin hypothesis })}{p(\text{ observed difference between samples} \mid \text{different origin hypothesis })} \tag{1}$$

Interpretation of a likelihood ratio

Coulthard & Johnson claim that a likelihood ratio of 8 does not tell us how much greater support there is for the same-origin hypothesis versus the different-origin hypothesis (p. 204), but in fact it does: A likelihood ratio of 8, assuming that it is accurate, gives exactly 8 times more support for the same-origin hypothesis than for the different-origin hypothesis, i.e., one would be 8 times more likely to obtain the measured differences between the samples if they had come from the same source than if they had come from different sources. Whatever the trier-of-fact's belief as to the origin of the questioned sample prior to the presentation of the forensic expert's evidence, after hearing it they should be 8 times more likely to believe that the origin of the questioned sample is the same as that of the known sample. (The trier-of-fact is the judge or the jury depending on the legal system.)

Combining evidence from multiple sources

Coulthard & Johnson state that the strength of evidence from multiple sources can be combined by multiplying the likelihood ratios together (p. 204–205). This is correct, but only if the sources of evidence can be assumed to be independent. The example they give is a multiplication of likelihood ratios obtained from comparisons of the frequency of occurrence of several different non-standard spellings. I would consider it unlikely that these would in fact be uncorrelated, e.g., I would expect that a high proportion of writers who substitute "their" for "there" also substitute "your" for "you're". For this example it would therefore not be appropriate to fuse the results by multiplying the likelihood ratios.

Semantically labelled interval scales

Coulthard & Johnson (pp. 205–206) claim that Rose [7] (pp. 61–62) advocates expressing the weight-of-evidence for ranges of likelihood ratio values using a five-interval scale with a semantic label associated with each interval, e.g., 1–10 limited support for the hypothesis, 10–100 moderate support, . . ., 10 000+ very strong support. This is not correct. Rose [7] (pp. 61–62) explains that such schemes have been proposed and discusses some of their pros and cons, but does not recommend either that they be used or not used. In fact, Rose does not support the use of such semantically labelled interval scales because they could have different meanings to different people [7] (p. 62), their definitions are circular [8] (p. 167), and they suffer from the cliff-edge effect [9].

Supplying the jury with a statistician

Coulthard & Johnson also suggest that Rose would advocate that juries be supplied with a statistician to assist them to determine priors and weigh all the evidence (p. 206, they refer to "an ideal Rosean world"). Rose has never advocated such a position. It is, however, similar to a proposal made by Lindley [10].

Admissibility

Coulthard & Johnson quote from the R v Adams appeal rulings ([1996] EWCA Crim 222, [1997] EWCA Crim 2474) that "the likelihood ratio '. . . was not appropriate for use in jury trials, nor as a means to assist the jury in their tasks.'" (pp. 206–207). Note that the words "likelihood ratio" were inserted here by Coulthard & Johnson, whereas the court actually referred to "Bayes' Theorem". "The calculation of the LR [likelihood ratio], however, is not a 'Bayesian analysis', as this term usually implies the assignment of prior probabilities" [4] (p. 200), and this difference is important.

If one only reads the two short quotes included in Coulthard & Johnson, then I believe one could easily arrive at an inadequate understanding of the court's rulings. In my (non-legal) opinion, a reading of the whole of the R v Adams appeal rulings indicates that what the court objected to was asking juries to determine a posterior probability of guilt via rigid application of mathematical formulae (such as Bayes' Theorem), and asking them to assign subjective numeric values to evidence which is not a result of scientific comparison (specifically non-DNA evidence in the R v Adams case). I would encourage the reader to consult the rulings for themself (available at http://www.bailii.org/), and also the discussions of the case by Balding [3] (pp. 149–151) and Donnelly [11], and comments in the R v GK appeal ruling ([2001] NSWCCA 413, available at http://beta.austlii.edu.au/).

I find no statement in the R v Adams appeal rulings which would indicate that the court disapproved of the use of likelihood ratios for the evaluation and presentation of scientific forensic comparison evidence. In the R v GK appeal the court ruled that the R v Adams rulings should not be interpreted so as to exclude the presentation of forensic evidence in the form of numeric statements of the probability of evidence given hypotheses (specifically in R v GK, likelihood ratios derived from DNA testing).

Further, in the R v Dohney & Adams appeal ([1996] EWCA Crim 728) the court effectively ruled that forensic scientists should use the likelihood-ratio framework. The court ruled that a forensic expert in DNA should provide "the frequency with which the matching DNA characteristics are likely to be found in the population", i.e., the match probability. The match probability is an alternative expression of a likelihood ratio which can be used in relation to DNA comparison evidence because of particular properties of DNA profiles. DNA profiles consist of discrete level values (e.g., counts of short tandem repeats known as alleles) from a finite number of measurements (each at a specific locus). If one discounts possibilities such as organ transplants and contamination, then the DNA profile of an individual organism does not change from occasion to occasion, and the probability of obtaining identical profiles under the same-origin hypothesis is 1, and the probability of obtaining non-identical profiles under the same-origin hypothesis is 0. The numerator of the likelihood ratio from a comparison of DNA profiles is therefore either 1 or 0 [2,12]. If the numerator is 0, then the denominator is irrelevant, the likelihood ratio is 0 and unless there has been an organ transplant, contamination, etc. then the samples do not have the same origin. If the numerator is 1, then the value of the likelihood ratio is determined by the denominator, the probability of finding an individual (other than the defendant) in the relevant population who has the DNA profile in question. The match probability is therefore simply the inverse of the likelihood ratio given in Equation 1, i.e., it is the probability of obtaining the same particular DNA profile in the questioned sample as in the known sample under the different origin versus the same origin hypothesis [13] (p. 484). Note that, unlike DNA profiles, data from forensic voice comparisons and forensic linguistic comparisons are typically continuous and subject to variability from occasion to occasion, and thus a match is not possible, numerators are never 0 or 1, and strength-of-evidence can be expressed using a full likelihood ratio but not via a match probability.

Prior odds

Coulthard & Johnson (p. 206) refer to Rose's [7] (pp. 63–64) discussion of prior odds, and the examples of a telephone call being made by one person where it is known that there were only five people or only two people in the house who could have made the telephone call. As mentioned above, they incorrectly claim that Rose advocates that a jury be supplied with a statistician to aid them to calculate prior odds. In fact Rose uses these examples simply to illustrate the effect of differing priors – his purpose is didactic not advocatory. His key point is actually that the forensic scientist should not and cannot calculate or know the prior odds and therefore should not and cannot legally or logically calculate posterior odds (this point is reiterated more forcefully in [8] p. 162). There are two interpretations of the rôle of priors in Bayes' Theorem:

(1) One interpretation is that the priors reflect the belief of the decision maker as to the probabilities of the competing hypotheses prior to them being presented with the evidence. Since such beliefs would be in the mind of the trier-of-fact at the time of the trial, the forensic scientist cannot know the trier-of-fact's prior beliefs.

(2) Another interpretation is pragmatic in that it assumes that one can determine reasonable priors using available information, as in the five-speakers-in-the-house example. In fact, scientists and engineers use such pragmatic priors all the time when building statistical models such as discriminant analysis models. Typical procedures are to give equal priors to every category group, or to weight the groups' priors according to the relative frequency of occurrence of tokens of each group in a training set. In the five-speakers-in-the-house example, Rose [7] (pp. 63–64) suggested that practical priors could be equal priors for each person in the house, i.e., a prior probability of 0.2 for each speaker or prior odds of $0.2/(1-0.2) = 1/4$. While I agree that in general these would be reasonable pragmatic priors, I disagree with Rose that these would be theoretically reasonable in a forensic situation. What if one of the speakers likes talking on the phone and another does not? What if one has laryngitis on that day? What if one is deaf? In general, there are too many unknown possibilities. However, Rose's discussion of pragmatic priors is theoretical, because even if a forensic scientist did have information available to allow them to calculate pragmatic priors, they should not, because in doing so they would be usurping the rôle of the trier-of-fact. In fact Rose argues that the forensic scientist should not calculate priors, and it is best if they know as little as possible about the circumstances of the crime so that there cannot even be any question that their assessment of the evidence could be influenced by anything other than a scientific evaluation of the samples provided to them. It is a particular strength of the likelihood-ratio framework that it is resistant to the forensic scientist's weight-of-evidence statement being influenced by their own personal beliefs or by any other evidence related to the case.

It is the interpretation in terms of the decision maker's beliefs that is probably most applicable in a legal trial. Although it may be inadvisable to ask the trier-of-fact to work with numeric expressions and formulae, and few people would naturally do this of their own accord, at the beginning of the trial the trier-of-fact must have some initial belief as to the guilt of the defendant. This is effectively a pragmatic prior which they have chosen for themself, perhaps influenced by a concept such as "innocent until proven guilty". Then, as each piece of evidence is presented, it will alter their belief so that it leans in one direction or the other, until at the end of the trial they must decide whether their belief in the guilt of the defendant exceeds some threshold such as "beyond a reasonable doubt" or "on the balance of probabilities".

Population samples

Finally, Coulthard & Johnson do raise one important area in which more work is required in order to bring about large-scale practical implementation of the likelihood-ratio framework for forensic-voice-comparison and forensic-linguistic-comparison casework (p. 205). In order to calculate likelihood ratios for a forensic voice comparison one must assess the similarity and the typicality of the known and questioned samples, and in order to do the latter one must have a relatively large number of voice samples from the potential population of offenders. If one is not to collect these on an ad-hoc case-by-case basis, then one must have access to a large database of voice samples from which to extract an appropriate subset for each case. Likewise, for a forensic linguistic comparison one would need a database of samples containing the relevant linguistic data.

Collecting and analysing such databases is potentially time consuming and expensive, and Coulthard & Johnson propose that we live "in a world where lawyers and courts are not willing to pay for what might be thought to be basic research." I do not share their pessimism about the potential for obtaining funds for the collection of databases for forensic application, in fact I believe that given the current concerns about crime and terrorism in many parts of the world this is a research activity for which it should be relatively easy to obtain funding from national governments and law-enforcement agencies. For example, the Australian Research Council (ARC) has specifically identified safeguarding Australia from terrorism, crime, and other threats as a priority area for research funding (my colleagues and I are currently preparing a major grant application in which we are seeking funding from the ARC and Australian law-enforcement agencies for forensic-voice-comparison research including the compilation of a database of 1000+ Australian English voices). Over the last quarter of a century, a great deal of time and money has been expended worldwide on collecting DNA databases for forensic use, and presentation of evidence from DNA comparison is now a ubiquitous component of criminal trials. The Congress-commissioned National Research Council's 2009 report on improving forensic science in the United States [14] finds that:

> "The development of scientific research, training, technology, and databases associated with DNA analysis have resulted from substantial and steady federal support for both academic research and programs employing techniques for DNA analysis. Similar support must be given to all credible forensic science disciplines if they are to achieve the degrees of reliability needed to serve the goals of justice." (p. S-9)

Over the last decade, the Guardia Civil in Spain has spent hundreds of thousands of Euros on forensic-voice-comparison research, including the collection of large databases of Spanish voices [15], making presentation of likelihood-ratio forensic-voice-comparison evidence commonplace in Spanish courts (in 2008 the Guardia Civil submitted 98 forensic-voice-comparison reports to the courts).

Over the last three decades, forensic DNA experts have dealt with issues related to choosing the relevant population including differences in population statistics across different racial groups, and the need to collect DNA databases for minority groups in which there may be less diversity in DNA profiles and thus greater probability of finding a random match than in the larger population [13]. The challenges should not, I think, be harder for forensic scientists working in other fields.

Conclusion

In conclusion, Coulthard & Johnson's [1] (pp. 203–207) portrayal of the likelihood-ratio framework contains a number of inaccuracies. I hope that I have been able to correct these so as to aid the reader in arriving at a better understanding of the framework.

Acknowledgments

The writing of this letter was financially supported by Australian Research Council Discovery Grant No. DP0774115. Thanks to the following for helpful comments on an earlier draft of this letter: Dr. Gary Edmund, School of Law, University of New South Wales. Dr. David Lucy, Department of Mathematics & Statistics, Lancaster University. Dr. Philp Rose, School of Language Studies, Australian National University.

References

1. Coulthard M, Johnson A. An introduction to forensic linguistics: language in evidence. London (UK): Routledge; 2007.
2. Aitken CGG, Taroni F. Statistics and the evaluation of forensic evidence for forensic scientist. 2nd ed. Chichester (UK): Wiley; 2004.
3. Balding DJ. Weight-of-evidence for forensic DNA profiles. Chichester (UK): Wiley; 2005.
4. Champod C, Meuwly D. The inference of identity in forensic speaker recognition. Speech Communication 2000;31:193–203.
5. González-Rodríguez J, Drygajlo A, Ramos-Castro D, García-Gomar M, Ortega-García J. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech and Language 2006;20:331–355.
6. Lucy D. Introduction to statistics for forensic scientists. Chichester (UK): Wiley; 2005.
7. Rose P. Forensic speaker identification. London (UK) & New York: Taylor and Francis; 2002.
8. Rose P. Technical forensic speaker recognition. Computer Speech and Language 2006;20:159–191.
9. Rose P, Morrison GS. A response to the UK position statement on forensic speaker comparison. International Journal of Speech, Language & the Law 2009; in press pending approval of minor

revisions.

10. Lindley DV. Probability and the law. The Statistician 1977;26:203–220.
11. Donnelly P. Appealing statistics. Significance 2005;2:46–48.
12. Evett IW. Towards a uniform framework for reporting opinions in forensic science case-work. Science & Justice 1998;38:198–202.
13. Foreman LA, Champod C, Evett IW, Lambert JA, Pope S. Interpreting DNA evidence: a review. International Statistics Journal 2003;71:473-473.
14. National Research Council. Strengthening forensic science in the United States: a path forward. Washington (DC): National Academies Press; 2009
15. Ramos D, González-Rodríguez J, González-Domínguez J, Lucena-Molina JJ. Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. In: [editors unknown] Proceedings of Interspeech 2008 Incorporating SST 2008. Interspeech 2008 Incorporating SST 2008; 22–26 September 2008; Brisbane (Australia). [place of publication unknown]: International Speech Communication Association; 2008. p. 1493–1496.