

Response to the Law Commission of England and Wales Consultation Paper No 190 “The Admissibility of Expert Evidence in Criminal Proceedings in England and Wales: A New Approach to the Determination of Evidentiary Reliability”

http://www.lawcom.gov.uk/expert_evidence.htm

Submitted by:

Geoffrey Stewart Morrison BSc MTS MA PhD

Research Associate, School of Language Studies, Australian National University

Visiting Fellow, School of Electrical Engineering & Telecommunications, University of New South Wales

e-mail: geoff.morrison@anu.edu.au

URLs: <http://geoff-morrison.net> <http://forensic-voice-comparison.net>

1 Introduction

I am an academic researcher conducting research on forensic voice comparison. I also perform some forensic casework at the behest of both prosecution and defence. Although I am not resident in the United Kingdom, I am a British Citizen. I am also interested in the proposals to amend admissibility requirements in England and Wales because their adoption in that jurisdiction could have an influence on other UK and international jurisdictions leading to them adopting similar standards.

I am in full agreement with the proposal that the admissibility of forensic evidence should be determined according to its reliability, and I am generally in agreement with the proposals set out in the Consultation Paper. My primary concern, however, is that insufficient consideration has been given as to the *measurement of the reliability of forensic analysis systems* which compare the quantifiable properties of samples of known and questioned origin (systems which compare DNA profiles, fingerprint marks, glass fragments, tool marks, voice recordings, etc.). A proper consideration of this issue requires consideration of what I and many others believe to be the correct framework for the evaluation of forensic comparison evidence, the *likelihood-ratio framework*. I begin by describing the likelihood-ratio framework, then proceed to describe a procedure for measuring the reliability of forensic systems which output likelihood ratios (this discussion is based closely on parts of Morrison, 2009b). This is followed by some comments on Proposals 1 and 2, and on accreditation. Some comments have relatively broad relevance, but others are specific to my area of expertise, *forensic voice comparison*.

2 The likelihood-ratio framework

The likelihood-ratio framework is recommended by forensics statisticians for the evaluation of forensic comparison evidence (Aitken & Stoney, 1991; Aitken & Taroni, 2004; Balding, 2005; Buckleton, 2005; Evett, 1990; Evett *et al.*, 2000; Lucy, 2005; Robertson & Vignaux, 1995), and has been adopted as standard for the evaluation of the forensic comparison of DNA profiles (Foreman *et al.*, 2003).

In the likelihood-ratio framework the task of the forensic scientist is to provide the court with a *weight-of-evidence* statement in answer to the question:

How much more likely are the observed differences/similarities between the known and questioned samples to arise under the hypothesis that they have the same origin than under the hypothesis that they have different origins?

The answer to this question is quantitatively expressed as a *likelihood ratio*, calculated using the schema given in Equation 1.

$$\text{LR} = \frac{p(E | H_{\text{so}})}{p(E | H_{\text{do}})} \quad (1)$$

Where LR is the likelihood ratio, E is the evidence, i.e., the measured differences between the samples of known and questioned origin, H_{so} is the same-origin hypothesis, and H_{do} is the different origin hypothesis. If the evidence is more likely to occur under the same-origin hypothesis than under the different-origin hypothesis then the value of the likelihood ratio will be greater than 1, and if the evidence is more likely to occur under the different-origin hypothesis than under the same-origin hypothesis then the value of the likelihood ratio will be less than 1. The size of the likelihood ratio is a numeric expression of the weight (or strength) of the evidence with respect to the competing hypotheses. If the forensic scientist testifies that one would be 100 times more likely to observe the differences between the known and questioned samples under the same-origin hypothesis than under the different-origin hypothesis (LR = 100), then whatever the trier of fact's belief prior to hearing this, they should now be 100 times more likely to believe that the samples have the same origin. Likewise, if the forensic scientist testifies that one would be 1000 times more likely to observe the evidence under the different-origin hypothesis than under the same-origin hypothesis (LR = 1/1000), then whatever the trier of fact's prior belief, they should now be 1000 times more likely to believe that the samples have different origins.

The numerator of the likelihood ratio can be considered a *similarity* term, and the denominator a *typicality* term. In calculating the weight of evidence, the forensic scientist must consider not only the degree of similarity between the samples, but also the degree of typicality of the samples with respect to the relevant population. Similarity alone does not lead to strong support for the same-origin hypothesis. For example, if two samples are determined to be very similar on an objective measure of some physical property, this is of little value if these physical properties are also very typical and samples selected at random from any two individuals in the relevant population are likely to be equally or more similar. On the other hand, if two samples are found to be very similar in terms of properties which are atypical in the population, then samples selected at random from any two individuals in the relevant population are unlikely to be equally or more similar. In general, more similarity and less typicality lead to greater support for the same-origin hypothesis, and less similarity and more typicality lead to greater support for the different-origin hypothesis. In order to calculate a quantitative estimate of typicality of the known and questioned samples, the forensic scientist must have access to a database of samples which are representative of the relevant population.

3 Why the forensic scientist must present the probability of evidence, and must not present the probability of hypotheses

A likelihood ratio is an expression of the probability of obtaining the evidence given same- versus different-origin hypotheses. There are logical and legal reasons why the forensic scientist must present a weight-of-evidence statement in this form and must not present the probability of the hypotheses given the evidence. Determining the probability of guilty versus not-guilty and whether this exceeds a threshold such as “beyond a reasonable doubt” or “on the balance of probabilities” is the task of the trier of fact. If the forensic scientist were to present the probability of same-origin versus different-origin and the evidence were potentially incriminatory, then they would be usurping the rôle of the trier of fact. The trier of fact does not make their decision on the basis of a single piece of evidence, rather their task is to come to a decision after having weighed all the evidence presented in court. What they require from a forensic scientist is a statement of the weight of a specific piece of evidence. One forensic scientist may present the weight of evidence related to specific DNA samples, another may present the weight of evidence related to specific fingerprint samples, etc., and the trier of fact will weigh all of these together. Not all the evidence presented will be scientific numeric evidence, and the trier of fact must consider both the weight of scientific numeric evidence and the weight of non-scientific non-numeric evidence. In addition, before any evidence has been presented the trier of fact will have some belief as to the guilt of the defendant, perhaps influenced by concepts such as “innocent

until proven guilty”, and this will also contribute to their final decision.

If a forensic scientist wanted to calculate the probability of same-origin versus different-origin hypotheses they would have to apply Bayes’ Theorem. The odds form of Bayes’ Theorem is provided in Equation 2.

$$\frac{p(H_{so} | E)}{p(H_{do} | E)} = \frac{p(E | H_{so})}{p(E | H_{do})} \times \frac{P(H_{so})}{P(H_{do})} \quad (2)$$

posterior odds
likelihood ratio
prior odds

In order to calculate the posterior odds, the forensic scientist would need to know the prior odds. Under one interpretation of Bayes’ Theorem, the prior odds would represent the trier of fact’s belief in the relative likelihood of the two hypotheses prior to the evidence being presented. Obviously the forensic scientist cannot know the trier of fact’s prior belief. Under another interpretation pragmatic priors can be calculated, e.g., if the crime were committed on an island and there are known to have been 100 people on the island at the time of the crime, then a pragmatic prior could be 1/100; however, this would involve the assumption that each person on the island is equally likely to have committed the crime, and although it may be appropriate for the trier of fact to make such an assumption, it is not appropriate for the forensic scientist to do so (and if other evidence has already been presented in the trial, it is unlikely that the trier of fact’s belief as to guilty versus non-guilty would still be 1/100 immediately prior to the presentation of the likelihood ratio from the forensic evidence in question). It is inappropriate for the forensic expert to present the posterior odds because the posterior odds include information and assumptions from sources other than an objective scientific evaluation of the known and questioned samples provided to them for evaluation. If the forensic scientist were allowed to present posterior odds then it would be possible that their testimony could be influenced by their own subjective conscious or unconscious opinion as to the guilt or innocence of the defendant. It is a strength of the likelihood-ratio framework that it is resistant to influence from such sources of bias. Note that the likelihood-ratio framework does not make use of prior probabilities and should not be confused with a full Bayesian framework (Buckleton, 2005; Champod & Meuwly, 2000; Rose, 2006).

4 The legal directive to present the probability of the evidence (the relationship between DNA match probabilities and likelihood ratios)

In *R v Doheny & Adams* ([1996] EWCA Crim 728) 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”. It may not be immediately obvious that this is a directive that forensic scientists should evaluate evidence using the likelihood-ratio framework; however, 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 differing 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 (Aitken & Taroni, 2004, p. 404; Evett, 1998). 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 DNA profile in the questioned sample as in the known sample under the different-origin versus the same-origin hypothesis (Foreman *et al.*, 2003, p. 484). Note that the discussion above is a simplification and some would argue that although at a molecular level DNA may be discrete, measurement includes non-discrete stages and the possibility of measurement error, hence at the analytical level the concept of “match” and likelihood-ratio numerators of 0 or 1 is not valid for DNA profiles, and the practice of reporting match probabilities should therefore be replaced with reporting of likelihood ratios. Measurements taken from most other objects of forensic examination (voice recordings, glass fragments, etc.) are inherently continuous, not discrete, and are subject to intrinsic variability such that multiple samples of the same object (multiple samples of the same speech sound produced by the same speaker, multiple fragments of glass from the same pane, etc.) are almost always measurably different, and forensic-comparison results can only be reported using likelihood ratios.

The *R v Adams* appeal rulings ([1996] EWCA Crim 222, [1997] EWCA Crim 2474) have been interpreted by some (e.g., Coulthard & Johnson, 2007) as prohibiting the presentation of likelihood ratios in court; however, as I argue in Morrison (in press), this is an incorrect interpretation (see also the discussion of *R v Adams* in the *R v GK* appeal ruling, [2001] NSWCCA 413). What the court objected to in *R v Adams* was not the use of the likelihood-ratio framework, but asking juries to determine a posterior probability of guilt via rigid application of mathematical formulae, and asking

them to assign subjective numeric values to evidence which is not a result of scientific comparison.

5 Measuring reliability

It is important to note that use of the likelihood-ratio framework does not guarantee reliability, rather it is a framework within which it is possible to measure reliability.

I refer below to a forensic comparison system, by which I mean the combination of the database of samples representative of the population of potential offenders (known as a background or reference database), the procedures and techniques for the extraction of numeric values which characterise the properties of the known and questioned samples and the samples in the database (measurement of the acoustic properties of voice recordings, measurement of the chemical composition of glass fragments, etc.), and the statistical procedures used to calculate likelihood ratios on the basis of these numeric values.

The reliability of a forensic comparison system can be assessed by testing it on a large number of pairs of samples where it is known whether each pair has the same origin or a different origin. Likelihood ratios greater than one favour the same-origin hypothesis and likelihood ratios less than one favour the different-origin hypothesis; however, forensic comparison of known and questioned samples is not a binary decision task but rather the task of determining the weight of evidence with respect to the same-origin versus different-origin hypotheses, i.e., the extent to which likelihood ratios are greater than or less than one, equivalently the extent to which log likelihood ratios are greater than or less than zero (the use of a logarithmic scale has the advantage of introducing symmetry between results supporting same- and different-origin hypotheses, LR values in the range zero to one are converted to log LR values in the range minus infinity to zero, and LR values in the range one to plus infinity are converted to log LR values in the range zero to plus infinity). Ideally a forensic comparison system would produce large positive log LR values for comparisons which are known to be same-origin comparisons and large negative log LR values for comparisons which are known to be different-origin comparisons. A system which produces a \log_{10} LR of 3 (LR = 1000) for a given same-origin comparison is better than a system which produces a \log_{10} LR of 1 (LR = 10) for the same same-origin comparison, and more importantly a system which produces a \log_{10} LR of -3 (LR = 1/1000) for a given same-origin comparison is worse than a system which produces a \log_{10} LR of -1 (LR = 1/10) for the same same-origin comparison (mutatis mutandis for different-origin comparisons). System reliability should not of course not be based on a single test measurement or a handful of test measurements, but rather on a global assessment of a large number of test measurements (the system which is more reliable overall could perform worse in some specific instances). A system which in tests produces a

large number of large-magnitude log LRs which support contrary-to-fact hypotheses would be unreliable and in danger of contributing to miscarriages of justice.

A metric which captures the gradient goodness of a set of likelihood ratios derived from test data is the log-likelihood-ratio cost, C_{llr} (Brümmer *et al.*, 2007; Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2007). C_{llr} has been adopted as a measure of system reliability both in forensic voice comparison (González-Rodríguez *et al.*, 2007; Morrison, 2009a; Morrison & Kinoshita, 2008; Ramos Castro, 2007) and in non-forensic automatic speaker recognition, including adoption by the US National Institute of Standards and Technology for use in its Speaker Recognition Evaluations (NIST SRE). I do not intend to suggest that C_{llr} specifically should be enshrined as a metric of reliability for forensic comparison systems, advances in thinking and research may leave some of its details obsolete, rather I present an outline of its calculation to provide a concrete example of what I would consider an appropriate procedure for measuring reliability.

Figure 1 provides a plot of a component of the C_{llr} function when it is known that the comparison is a same-origin comparison. Large positive log LR values which correctly support the same-origin hypothesis are assigned very low C_{llr} component values, log LR values close to zero provide little support for either the same-origin or different-origin hypotheses and are assigned moderate C_{llr} component values, and negative log LR values which contrary-to-fact support the different-origin hypothesis are assigned high C_{llr} component values which increase rapidly as the log LR values become more negative and provide stronger contrary-to-fact support for the different-origin hypothesis. For calculating a component of the C_{llr} function when it is known that the comparison is a different-origin comparison the plot would be a mirror of Figure 1, mirrored at $\log_{10} \text{LR} = 0$. The C_{llr} for the system given a test set of a large number of same-origin and different-origin comparisons is the balanced mean of all the same-origin C_{llr} component values and all the different-origin C_{llr} component values. The lower the C_{llr} , the better the performance of the system. If several systems are tested using the same set of test data, then the most reliable system is the system which results in the lowest C_{llr} value.

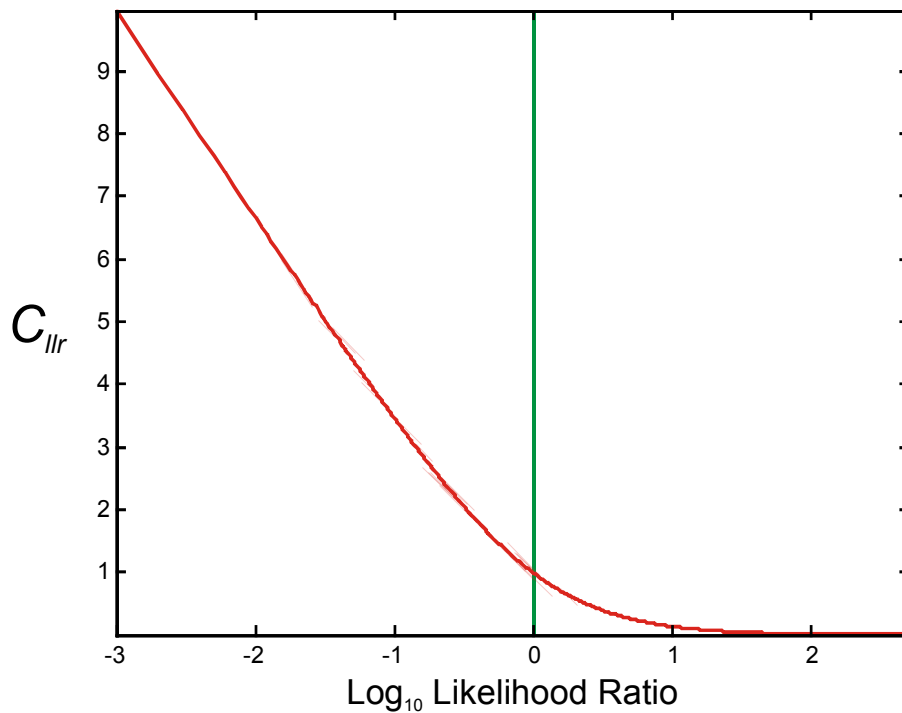


Figure 1. Component of the C_{lr} function when it is known that the likelihood ratio is the result of a same-origin comparison.

I believe that it is incumbent on forensic scientists to demonstrate the reliability of their forensic comparison systems using a metric such as C_{lr} . General system reliability can be demonstrated using test data and where appropriate the results submitted to peer-reviewed journals. It is important to note that such measures are based on the systems' performance on a particular test set, and if they are to be meaningful in a particular case the test set must be representative of the relevant population in that case. Reliability in a particular case should be demonstrated by ensuring that testing is conducted using background and test databases which are appropriate to the case, and extracting numeric values for the same features using the same procedures as applied to the known and questioned samples in the case. Such measures of system reliability should be included in the case report.

In a particular case, it is also possible to provide the court with an error rate for the specific likelihood ratio obtained from the comparison of the known and questioned samples, e.g., if a likelihood ratio of 10 in favour of the same-origin hypothesis is obtained, an error rate can be reported as the proportion of different-origin comparisons in a test set which resulted in likelihood ratios of equal to or greater than 10. It is possible that a forensic comparison system could be shown to be reliable overall and thus in principle admissible, but in a particular case could produce a likelihood ratio for the known versus questioned sample comparison which is close to one and has a high error rate, and thus provides little information of value to the trier of fact (except to cast doubt on strong assertions made by either the prosecution or defence).

6 Comments on Proposal 1

I fully endorse the view that forensic analyses which are more objective and whose reliability can be quantitatively demonstrated should be preferred over more subjective analyses for which it is harder to quantify reliability (4.49, 4.53, 4.55, 4.56, 6.10, 6.11, 6.14, 6.15, 6.18–6.20). I also fully endorse the view that in determining admissibility, demonstrable reliability of forensic analyses should take precedence over their degree of acceptance in the field, publication in refereed journals, or the reputation of the scientist presenting the evidence, etc. (4.28–4.31, 4.35, 4.36, 4.38, 4.49, 4.62). I therefore believe that demonstrable reliability should be the primary consideration for determining admissibility, and 6.26(1)(a) should take precedence over all other considerations, in particular it should be given much more weight than the sum of 6.26(1)(c)–(f).

Forensic voice comparison is a field in which there have been a number of approaches, some of questionable reliability. It may be that much of the forensic-voice-comparison evidence which has been admitted in courts in England and Wales in the past has not been tested and shown to be reliable. There has been considerable resistance from some forensic speech scientists to the adoption of the likelihood-ratio framework (see Morrison, 2009b, for a history of the adoption of the likelihood-ratio framework for forensic voice comparison). A large group of forensic speech scientist in the United Kingdom recently proposed a framework for the evaluation of forensic-voice-comparison evidence (French & Harrison, 2007), which is inconsistent with the likelihood-ratio framework, which, as far as I am aware, has not been tested and found to be reliable, and which I therefore believe would not meet the proposed admissibility requirements for scientific evidence. A full critique of French & Harrison (2007) is provided in Rose & Morrison (in press).

It may be that if Proposal 1 or similar becomes law, some forensic-voice-comparison practitioners could seek to have analyses of untested reliability admitted under the laxer rules for *experience-based expert evidence* (6.34–6.43) rather than the stricter rules for *scientific expert evidence* (6.25–6.33). I believe that in any branch of forensic science where there exist systems which are capable of meeting the requirements of the scientific category, testimony tendered under the experience-based category should be ruled inadmissible. I also believe that in forensic voice comparison, systems with the potential for meeting the requirements of the scientific category have been and are being developed (e.g., González-Rodríguez *et al.*, 2006; González-Rodríguez *et al.*, 2007; Morrison, 2009a). I also believe that systems of testable reliability could probably be developed for analysis of other types of evidence including ear-prints (6.33), and, counter the claim in 6.34 that their analysis would be “(non-scientific) experienced-based”, handwriting (6.34) and document examination (6.36). If the relevant properties of samples can be quantified then likelihood-ratio analyses can be applied.

González-Rodríguez *et al.* (2005), for example, show how to apply the likelihood-ratio framework to fingerprints, faces, and signatures.

7 Comments on Proposal 2

6.54(1) expresses the opinion that any objection to “the theory that each human being is endowed with a unique set of fingerprints” “would be regarded as manifestly unfounded and a pre-trial investigation into the validity of the theory or methodology would be unnecessary”. The proposition that this theory and any methodology based on it can be relied upon to provide forensic individualisation is, in fact, untenable. Even if the object of investigation is highly complex allowing for multiple features to be measured so as to characterise its properties, given a large enough database of samples for comparison it will be possible to find by chance a sample which is indistinguishable from the questioned sample. This is true for DNA profiles, fingerprint marks, and any other object of forensic comparison. Increasing the number of features measured and compared, or increasing the precision of the measurement techniques, will reduce the probability of finding a measurably indistinguishable match, but the probability of finding a match by chance will always be non-zero. Even ignoring the limitations of measurability, if we take DNA, which may be considered discrete at a molecular level, although the probability of finding two unrelated individuals with identical genomes is extremely small it is not zero. Fingerprint analysis empirically does not lead to forensic individualisation, as demonstrated by the case of Portland Oregon lawyer Brandon Mayfield who, on the basis of fingerprint evidence, was incorrectly identified by the FBI as being involved in the 2004 Madrid train bombings (see Saks & Kohler, 2005). Also, as Saks & Kohler (2005) point out, the issue of whether this error was due to inadequacies of human performance, measurement error, system performance, or theory is irrelevant, the reliability of a forensic analysis system should be demonstrated in practice, this is what matters. I also add that as explained above, the task of the forensic scientist should be to provide a weight-of-evidence statement in the form of a likelihood ratio, and it is not appropriate for them to give a statement of individualisation. A likelihood ratio is a probabilistic statement and even a likelihood ratio which provides very strong support for one hypothesis can be countered by strong contradictory evidence. An unquestioning belief in the infallible ability of forensic fingerprint analysis to provide individualisation leads to strong contradictory evidence being ignored, and in the case of the FBI and Mr. Mayfield the arrest of an innocent man.

8 Accreditation

I have reservations about the establishment of accreditation for forensic-voice-comparison

practitioners at the present time. I believe that given the present state of this branch of forensic science there is a danger that the implementation of such a system could entrench existing subjective unreliable practices and inhibit the use of newer more objective demonstrably reliable practices. I believe this because those who have adopted the more objective demonstrably reliable approaches to forensic voice comparison are still in the minority, and the development of such a system could be dominated by the self-preservation interests of the majority.

I also fear that if an accreditation system were to be established, then it, rather than a case-by-case examination of the reliability of tendered scientific testimony, could become the de facto gate-keeper for admissibility.

9 Education for judges and lawyers, and the appointment of court-appointed assessors

I agree with the idea that judges should be provided with education and be able to call upon court-appointed assessors to assist them in making better-informed decisions with respect to the admissibility of scientific expert evidence.

References

- Aitken, C. G. G., & Stoney, D. A. (1991). *The Use of Statistics in Forensic Science*. Chichester, UK: Horwood.
- Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist*. 2nd ed. Chichester, UK: Wiley.
- Balding D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. Chichester, UK: Wiley.
- Brümmer, N., Burget, L., Cernocký, J. H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio, Speech and Language Processing*, 15, 2072–2084.
- Brümmer, N., & du Preez, J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language*, 20, 230–275.
- Buckleton, J. (2005). A framework for interpreting evidence. In Buckleton, J., Triggs, C. M., & Walsh, S. J. (Eds.), *Forensic DNA Evidence Interpretation*, pp. 27–63. Boca Raton, FL: CRC.
- Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, 31, 193-203.
- Coulthard M, & Johnson A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. London, UK: Routledge.

- Evett, I. W. (1990). The theory of interpreting scientific transfer evidence. *Forensic Science Progress*, 4, 141–179.
- Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science case-work. *Science & Justice*, 38, 198–202.
- Evett, I. W., Jackson, G., Lambert, J. A., & McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements, *Science & Justice*, 40, 233–239.
- Foreman, L. A., Champod, C., Evett, I. W., Lambert, J. A., & Pope, S. (2003). Interpreting DNA evidence: A review. *International Statistics Journal*, 71, 473–473.
- French, J. P., & Harrison, P. (2007). Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law*, 14, 137–144.
- González-Rodríguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., & Ortega-García, J. (2005). Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International*, 155, 126–140.
- González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., & Ortega-García, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20, 331–355.
- González-Rodríguez, J., Rose, P., Ramos, D., Torre, D., & Ortega-García, J. (2007.) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15, 2104–2115.
- Lucy D. (2005). *Introduction to Statistics for Forensic Scientists*. Chichester, UK: Wiley.
- Morrison, G. S. (2009a). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125, 2387–2397.
- Morrison, G. S. (2009b). The place of forensic voice comparison in the ongoing paradigm shift. Invited paper to be presented at the *2nd International Conference on Evidence Law and Forensic Science*, Beijing, China, July 25–26. [Version also submitted to *Science & Justice* in May 2009. Pre-publication version available <http://forensic-voice-comparison.net>]
- Morrison, G. S. (In Press). Comments on Coulthard & Johnson’s (2007) portrayal of the likelihood-ratio framework. *Australian Journal of Forensic Sciences*, 41. [Pre-publication version available <http://forensic-voice-comparison.net>]
- Morrison, G. S., & Kinoshita, Y. (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. *Proceedings of Interspeech*

- 2008 *Incorporating SST 2008* (pp. 1501–1504). International Speech Communication Association.
- Ramos Castro, D. (2007). Forensic evaluation of the evidence using automatic speaker recognition systems. PhD dissertation, Universidad Autónoma de Madrid, Madrid, Spain.
- Robertson, B., & Vignaux, G. A. (1995) *Interpreting Evidence*. Chichester, UK: Wiley.
- Rose P. (2006). Technical forensic speaker recognition. *Computer Speech and Language*, 20, 159–191.
- Rose, P., & Morrison, G. S. (In Press). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law*, 16. [Pre-publication version available <http://forensic-voice-comparison.net>]
- Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, 309, 892–895.
- van Leeuwen, D. A., & Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In Müller C. (Ed.), *Speaker Classification I: Selected Projects*, pp. 330–353. Heidelberg, Germany: Springer-Verlag.