THE PLACE OF FORENSIC VOICE COMPARISON IN THE ONGOING PARADIGM SHIFT

Geoffrey Stewart Morrison[1,2]

[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
geoff.morrison@anu.edu.au

Abstract

We are in the midst of what Saks & Koehler (2005) have called a *paradigm shift* in the evaluation of evidence in the forensic comparison sciences. This is a shift to requiring that the evaluation of forensic evidence actually be scientific, including that the reliability of methodologies be testable, and requiring that forensic evidence be evaluated and presented to the courts in a logically correct manner. Reliability was a primary concern in the US Supreme Court's *Daubert* decision in 1993. The logically correct evaluation and presentation of the weight of forensic evidence was a primary concern in a number of court cases in the United Kingdom including the Appeal Court of England and Wales' 1996 ruling on the presentation of DNA evidence in *R v Doheny & Adams*. In the US National Research Council's report to Congress released in February 2009 current practice in the evaluation of nuclear DNA evidence is held up as a model to emulate, and current practices in other branches of forensic science are subject to sometimes severe criticism. In the present paper I examine the place of forensic voice comparison in the ongoing paradigm shift. Over the last decade a small number of forensic-voice-comparison researchers have been working in the post-shift paradigm. They have adopted the likelihood-ratio framework for the evaluation of forensic evidence, the same framework as used in DNA analysis. I provide a brief description of the likelihood-ratio framework, followed by a brief history of the adoption of the likelihood-ratio framework for forensic voice comparison by the research community, and by the forensic practitioner, law-enforcement, and judicial communities.

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

This is a written version of an invited presentation given at the 2nd International Conference on Evidence Law and Forensic Science, Beijing, China, 25–26 July, 2009. It is slightly revised from the version which appeared as:

Morrison, G. S., (2009). The place of forensic voice comparison in the ongoing paradigm shift. *The 2nd International Conference on Evidence Law and Forensic Science Conference Thesis* (Vol. 1, pp. 20–34). Beijing, China: The Key Laboratory of Evidence Science of the Ministry of Education (The Institute of Evidence Law and Forensic Science, China University of Political Science and Law).

1 THE NEW PARADIGM IN FORENSIC SCIENCE

1.1 The ongoing paradigm shift

Today we are in the midst of what Saks & Koehler [1] have called a *paradigm shift* in the evaluation and presentation of evidence in the forensic sciences which deal with the comparison of the quantifiable properties of samples of known and questioned origin, e.g., DNA profiles,

fingerprints, bullets, glass fragments, handwriting, and voice recordings. This is a shift towards requiring that the forensic evaluation of evidence be based on scientific methodologies which can be demonstrated to produce reliable results. It is also a shift towards requiring that forensic evidence be evaluated and presented in a logically correct manner which will not unfairly favour either the prosecution or the defence.

Seminal legal events in the paradigm shift occurred in the 1990s: In *Daubert v Merrell Dow Pharmaceuticals* (92-102) 509 US 579 [1993] the US Supreme Court ruled that when considering the admissibility of scientific evidence, the judge must consider whether the methodology is scientifically valid and whether it is reliable, i.e., whether it has been empirically tested, whether it is replicable, and whether it has a known and acceptable error rate. In *R v Doheny & Adams* [1996] EWCA Crim 728 the Appeal Court of England and Wales ruled that a forensic expert must present the probability of observing the evidence given the hypotheses of same versus different origin.

Also in the 1990s the relatively new field of forensic DNA comparison was rapidly developing and assuming a central role in many criminal cases. Arguably the newness of forensic DNA comparison, and the strong scientific training of those developing it, made this branch of forensic science ideal for the implementation of a statistical framework for the evaluation of evidence which would allow it to meet the requirements of the *Daubert* and *Doheny & Adams* rulings.

Despite more than a decade having past since these events, the propagation of the paradigm shift to other branches of forensic science has been slow. In February 2009 the US National Research Council (NRC) released its report to Congress on strengthening forensic science in the United States [2]. The report concluded that:

> "[S]ome forensic disciplines are supported by little rigorous systematic research to validate the discipline's basic premises and techniques." (S-16)

> "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." (S-9)

1.2 The likelihood-ratio framework

The framework which in the 1990s was adopted as standard for the forensic comparison of DNA profiles is the *likelihood-ratio framework* (see Foreman *et al.* [3] for a history of statistical procedures applied to the evaluation of DNA evidence, and Evett [4] for developments in forensic statistics immediately prior to the advent of forensic DNA analysis). Descriptions of the likelihood-ratio framework can be found in several textbooks and articles including [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. 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.

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

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 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 potential population of offenders. 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 very 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 leads to greater support for the same-origin hypothesis, and less similarity and more typicality leads 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 potential population of offenders.

1.3 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.

$$\frac{p(\mathrm{H_{so}} \mid \mathrm{E})}{p(\mathrm{H_{do}} \mid \mathrm{E})} = \frac{p(\mathrm{E} \mid \mathrm{H_{so}})}{p(\mathrm{E} \mid \mathrm{H_{do}})} \times \frac{P(\mathrm{H_{so}})}{P(\mathrm{H_{do}})} \qquad (2)$$

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 presented 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 (avoidance of human bias is a major concern in the NRC report, see Recommendation 5). Note that the likelihood-ratio framework does not make use of prior probabilities and should not be confused with a full Bayesian framework [7, 8, 14].

1.4 Comparison not identification

An important terminological point which arises from the discussion above, is that the forensic scientist does not (should not) "identify" the suspect, because "identification" implies determining a posterior probability. The forensic sciences such as DNA analysis and fingerprinting which have traditionally been referred to as "identification sciences" should not therefore be referred to as such. The term "comparison" is more appropriate [15]. Unfortunately, the authors of the NRC report do not appear to have appreciated this and continually refer to "identification" and "individualization". I fear that this is not purely a terminological issue but that it may reflect a failure to appreciate the underlying concepts (see Meuwly [16] on terminological and logical problems with the use of these terms in forensic science). I see it as a weakness of the report that while it holds up nuclear DNA analysis as a model to emulate in general, it does not discuss the theoretical and methodological aspects of the statistical evaluation of DNA evidence which should be emulated. The term "likelihood ratio" appears only once in the report, and this is in the title of a cited paper; however, the report does recommend Aitken & Taroni [5], Evett [4], and Evett *et al.* [17] as providing "the essential building blocks for the proper assessment and communication of forensic findings"(6-3), and all three advocate the use of the likelihood-ratio framework.

A terminological aside: Following Meuwly's [16] logic we also should actually be using a term such as "forensic comparison of voice recordings" rather than "forensic voice comparison", i.e., it is the properties of the recordings which are actually compared, not the voices themselves, and certainly not the speakers. Since the "of" construction has the potential to interfere with the understanding of sentence structure, I will continue to use the somewhat less exact "forensic voice comparison".

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

In *Doheny & Adams* 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 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 [5 p. 404, 18]. 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 [3 p. 484]. Note that the discussion

above is a simplification and some would argue that at the analytical level DNA data are not discrete, that the concept of "match" and likelihood-ratio numerators of 0 or 1 is therefore not valid for DNA profiles, and that the practice of reporting match probabilities should therefore be replaced with reporting of likelihood-ratios (personal communication from Didier Meuwly, April 2009).

1.6 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. 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. Unfortunately, the authors of the NRC report appear to have failed to appreciate this and their discussion of error rates is framed in terms of correct versus incorrect identifications and exclusions (4-5–4-9). 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} [19]. The lower the C_{llr}, the better the performance of the system. C_{llr} has been adopted in the US National Institute of Standards and Technology Speaker Recognition Evaluations (NIST SRE) as a measure of the reliability of the system in general, and is extremely useful in research; however, in a particular court case an error rate for the specific likelihood ratio obtained can also be reported, e.g., if a likelihood ratio of 100 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 100. It is important to note that such measures are based on the system's 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 potential population of offenders in that case.

2 FORENSIC VOICE COMPARISON AND ITS PLACE IN THE PARADIGM SHIFT

2.1 Differences between voices and DNA

Data obtained from measurements of the acoustic properties of human voices are very different from DNA profiles. Acoustic data are continuous not discrete, and a speaker never says the same thing exactly the same way twice, so there is always variation from occasion to occasion. It is therefore impossible to obtain a "match" between two voice samples, indeed the concept is meaningless with respect to continuously valued data. Because of this, the strength of evidence from a forensic voice comparison cannot be expressed as a match probability, and must be expressed in the form of a full likelihood ratio. Acoustic properties which have relatively small within-speaker variance and relatively high between-speaker variance will be useful for forensic voice comparison in that they will potentially lead to likelihood ratios which deviate from 1. A typical forensic-voice-comparison system would be based on measurements of a number of acoustic properties. Although forensic DNA comparison can routinely result in likelihood ratios in the tens of millions, expectations of potential maximum values in the hundreds or thousands may be more reasonable for forensic voice comparison.

2.2 Approaches to forensic voice comparison

Historically it is possible to identify at least four different approaches to forensic voice comparison: *auditory*, *spectrographic*, *acoustic-phonetic*, and *automatic*. For simplicity of exposition these will be treated as discrete, but in practice it has not been uncommon for aspects of two approaches to be combined, e.g., auditory-spectrographic and auditory-acoustic-phonetic. The descriptions of each approach below are meant to be rough sketches rather than thorough expositions (for fuller descriptions of the first three approaches see Rose [13], and for the latter see Bimbot *et al*. [20] and Ramos Castro [21]).

2.2.1 The auditory approach

The auditory approach is practised by phoneticians who may be drawing on years of training and experience in auditory phonetics, a tradition which includes using phonetic symbols and diacritics to transcribe the speech sounds which are heard. The phoneticians listen to the known and questioned voice samples and comment on any properties of the voices which may be shared and which in their experience they consider unusual, distinctive, or otherwise noteworthy, or any features which are noteworthy because they are present in one sample and unexpectedly absent in another. Conclusions have typically been in the form of qualitative expressions of certainly that the samples have either the same or different origin (subjective posterior probabilities). Although theoretically it would be possible to obtain a measure of reliability by having a phonetician compare a large number of pairs of samples known by the tester (but not by the testee) to be of same or different origin, as far as I am aware no large-scale tests of this approach have been conducted. Although it is possible that an experienced practitioner of the auditory approach could be proven to have a high degree of reliability, the approach would still fail to meet tests of scientific rigour. For example, since the approach is based on subjective judgment rather than objective measurement, it would be impossible to devise a precise set of instructions which would allow independent replication of the methodology.

2.2.2 The spectrographic approach

The spectrographic approach, also known as *voiceprinting*, is based on a technology developed in the 1940s which allows the time-varying amplitude of the frequency properties of an acoustic signal to be visually displayed. Typically time is on the *x*-axis, frequency on the *y*-axis, and amplitude within this two-dimensional graph is represented by the darkness of a monochrome scale, see Figure 1. To the layperson the conversion from the acoustic domain to the visual domain may give the impression that the approach is scientific, but in fact the analysis is not objective, it consists of the practitioner visually comparing a number of spectrograms in order to arrive at a qualitative expression of the probability of same or different origin (subjective posterior probabilities). Although it has some diehard supporters, the general conclusion of the scientific community is that the spectrographic approach is not scientific and not reliable – for summaries see Gruber & Poza [22], Rose [13 pp. 107–122] and, from a legal perspective, Solan & Tiersma [23]. In July 2007, a meeting of the International Association for Forensic Phonetics and Acoustics (IAFPA) passed a resolution that:

> "The Association considers this approach to be without scientific foundation, and it should not be used in forensic casework." <http://www.iafpa.net/voiceprintsres.htm>

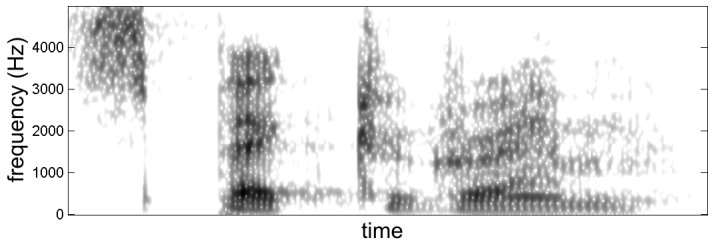

Figure 1. Example of a spectrogram.

2.2.3 The acoustic-phonetic approach

The acoustic-phonetic approach, as its name suggests, was developed by phoneticians trained in acoustic phonetics and involves making objective measurements of the acoustic properties of speech sounds. Comparable phonetic units are identified in both known and questioned voice samples and then acoustic properties of these units are measured. An example of a phonetic unit is the vowel /aɪ/ (the vowel sound in the words "I", "hi", "buy", etc.). A phonetic unit could be a phoneme (a basic building block of speech in phonological theory), but could also cover a shorter or longer stretch of speech. Examples of acoustic properties are the resonances of the vocal tract (formants) which in phonetic theory are primary acoustic correlates of vowel category (phoneme) identity, i.e., they are the primary cues used by listeners to determine whether a speaker said /aɪ/, /aʊ/ (the vowel sound in "now", "cow", etc.), /æ/ (the vowel sound in "hat", "cap", etc.), etc.. The time and expense involved in data analysis is a major constraint on the application of the acoustic-phonetic approach: Identification of phonetic units and measurement of their acoustic properties is a labour-intensive manual or semi-automated process which must be conducted by someone with expertise in acoustic phonetics.

2.2.4 The automatic approach

The automatic approach was developed by signal-processing engineers. As with the acoustic-phonetic approach, it is based on objective measures of the acoustic properties of speech, but typically no attempt is made to exploit information relating to phonetic units. Typical features in an automatic system are short-term spectra (20–30 ms) extracted over the entire duration of the speech recording and quantified using cepstral coefficients (an explanation of these features accessible to a lay audience would be rather involved and is not warranted here). Although a typical automatic system treats fine-grained phonetic information as noise (unwanted variability), it has the major advantage of being able to rapidly and cheaply process massive amounts of data.

Of these four approaches to forensic voice comparison only the latter two, acoustic-phonetic and automatic, can and have been implemented in such a way as to provide quantitative likelihood-ratio output of measurable reliability.

2.3 The adoption of the likelihood-ratio framework by the research community

2.3.1 Proposals to adopt the likelihood-ratio framework

The first published proposal for the adoption of the likelihood-ratio framework for forensic voice comparison appears to have been made by S. R. Lewis in 1984 [24]. This clearly had very little effect on the research community because there was then a decade-long hiatus before the idea appeared in publication again. At the *International Congress of Phonetic Sciences* (ICPhS) in August 1995 A.

P. A. Broeders briefly stated that forensic voice comparison evidence should be evaluated using likelihood ratios [25]. In articles published in Australian journals in 1997, 1999, and 2001 Philip Rose also proposed that forensic voice comparison evidence should be evaluated using likelihood ratios [26, 27, 28]. Rose cites Robertson & Vignaux [12], recommended to him by Hugh Selby, as a formative influence (personal communication from Philip Rose, April 2009). A more substantial argument which has had a greater impact on the research community was made by Christophe Champod and Didier Meuwly, initially at the (*Reconnaissance de Locuteur et ses Applications Commerciales et Criminalistiques*) *RLA2C Workshop* in April 1998, with a subsequent version submitted as a journal article in October 1998 and published in *Speech Communication* in September 2000 [29, 8]. This paper drew on the existing literature on the evaluation and interpretation of forensic evidence in fields such as DNA to make a lucid argument for its adoption in forensic voice comparison. Meuwly cites Kwan [30], Lewis [14], and Evett & Buckleton [31] as formative influences (personal communication from Didier Meuwly, April 2009).

Didier Meuwly and Andrzej Drygajlo also described the application of the likelihood-ratio framework to forensic voice comparison at the *Congres Français d'Acoustique* in September 2000 [32]. In December 2001 at the *International Conference on Law and Language – Prospect and Retrospect* Francis Nolan suggested the use of the likelihood-ratio framework as a conceptual framework for acoustic-phonetic forensic voice comparison, but expressed doubts as to the practicality of the quantitative implementation of the framework [33]. At two successive *Interpol Forensic Science Symposia*, in 2001 and 2004, A. P. A. Broeders presented reviews of developments in forensic voice comparison from 1998 to 2001 and 2001 to 2004 respectively [34, 35]. In both reports he discussed the need for forensic voice comparison evidence to be evaluated using the likelihood-ratio framework, and noted that a number of automatic systems could output likelihood ratios.

2.3.2 Implementation of the likelihood-ratio framework in automatic forensic voice comparison

The first automatic systems specifically designed to output likelihood ratios for forensic application were developed by a research group working in Lausanne, Switzerland, and a couple of years later they were followed by a research group working in Madrid, Spain. In April 1998 Didier Meuwly, Mounir El-Maliki, and Andrzej Drygajlo of the Lausanne group presented a paper at the (*Continuous Speech Recognition Over the Telephone*) *COST-250 Workshop*. They described the rationale for the use of the likelihood-ratio framework for forensic voice comparison, and described the design and results of tests of a Gaussian-Mixture-Model (GMM) system that calculated likelihood ratios [36]. The paper was not well received, one audience member described the likelihood-ratio framework as nonsense. Articles which the group submitted to journals were also rejected because of a lack of understanding of the framework on the part of the reviewers (personal communication from Didier Meuwly, April 2009). This situation was soon to change: At the *RLA2C Workshop* in April 1998, Session Chair George Doddington recommended the use of the likelihood-ratio framework. At the *International Speech Communication Association* (ISCA) *Odyssey Speaker Recognition Workshop* in June 2001, papers describing likelihood-ratio GMM automatic forensic-voice-comparison systems were presented by Anrzej Drygajlo and Didier Meuwly of the Lausanne group, and by Joaquín González-Rodríguez, Javier Ortega-García, and José Juan Lucena-Molina of the Madrid group [37, 38]. Didier Meuwly's PhD dissertation was also completed in 1999 and published in 2001 [39].

Since then, the likelihood-ratio framework has gradually become established as standard within the automatic forensic-voice-comparison research community. The Forensic Speaker Recognition Evaluation conducted in the fall of 2003 by the Netherlands Forensic Institute and the Netherlands Organization for Applied Scientific Research (NIF-TNO) included evaluation of likelihood-ratio results [40], and (although their goal is not primarily forensic) evaluation via likelihood ratios (in the form of C_{llr}) was adopted by the NIST SRE in 2006.

Important journal articles describing the likelihood-ratio framework and its use for the robust calculation of likelihood ratios in automatic forensic voice comparison were described in several journal articles published by the Lausanne and Madrid groups in the middle of the decade [41, 42, 9, 43, 10].

At ISCA's *Interspeech* conference in September 2008, a keynote address was given by Joaquín González-Rodríguez in which the likelihood-ratio framework was a central focus. Also at *Interspeech 2008* a tutorial on likelihood-ratio forensic voice comparison (both automatic and acoustic-phonetic) was presented by Yuko Kinoshita, Geoffrey Stewart Morrison (both members of the Canberra group, see section 2.3.3), and Daniel Ramos (a member of the Madrid group).

2.3.3 Implementation of the likelihood-ratio framework in acoustic-phonetic forensic voice comparison

In acoustic-phonetic forensic voice comparison the adoption of the likelihood-ratio framework has been pioneered by a research group working in Canberra, Australia (lead by Philip Rose). The first implementation of likelihood-ratio acoustic-phonetic forensic voice comparison was in Yuko Kinoshita's PhD dissertation completed in 2001 [44]. In 2002 and 2003 Philip Rose published a book and a book-length chapter on likelihood-ratio forensic voice comparison, one aimed primarily at phoneticians [13], and the other aimed primarily at the legal community [45]. Although now somewhat dated, Rose [13] has become a standard reference for likelihood-ratio acoustic-phonetic forensic voice comparison.

Additional expositions on the use of the likelihood-ratio framework for acoustic-phonetic forensic voice comparison authored by Philip Rose were published in journal articles in the middle of the decade [46, 14], and journal articles by the Canberra group reporting research results obtained using the framework include Rose *et al*. [47], Kinoshita [48], and Morrison [49, 50]. A survey of forensic phonetics authored by Michael Jessen of the German Federal Police was published in 2008. Within this Jessen recommended the adoption of the likelihood-ratio framework [51]. There is, however, ongoing resistance to the adoption of the likelihood-ratio framework from a large part of the forensic-phonetics community (see section 2.5).

Recently Cuiling Zhang of the China Criminal Police University in Shenyang has collaborated with the Canberra group producing the first applications of the likelihood-ratio framework to Chinese speech [52, 53].

2.3.4 Combination of automatic and acoustic-phonetic approaches within the likelihood-ratio framework

There is increasing interest in combining aspects of the automatic and acoustic-phonetic approaches to forensic voice comparison within the likelihood-ratio framework. Philip Rose is currently leading a research project on this topic funded by the Australian Research Council from 2007 to 2009 This includes collaboration with the Madrid group and with a group at the University of New South Wales in Sydney, Australia, which began working on forensic voice comparison in 2007 (the Sydney group is lead by Eliathamby Ambikairajah). Another project investigating

automatic and acoustic-phonetic approaches to forensic voice comparison (lead by Michael Jessen) is a collaboration between the German Federal Police, the Romanian Ministry of Justice, and the Austrian Academy of Science, funded by the European Union from 2008 to 2010. Also, a special session on combining automatic and acoustic-phonetic approaches was organised by Geoffrey Stewart Morrison at *Interspeech 2008*, and included papers from the Canberra, EU, Madrid, and Sydney groups. Journal articles with combinations of acoustic-phonetic and automatic techniques include [10, 50].

2.4 The adoption of the likelihood-ratio framework by the forensic practitioner, law-enforcement, and judicial communities

2.4.1 Spain

The only jurisdiction were forensic voice comparison can be said to be commonly practised using the likelihood-ratio framework is Spain. In 1997 the Guardia Civil began funding research to develop an automatic forensic-voice-comparison system, and in 2004 they began creating a large database of Spanish voices. The research was conducted by the Madrid group, which was initially based at the Polytechnic University of Madrid but moved to the Autonomous University of Madrid in 2005. By 2005 the system, which is called IdentiVox, produced likelihood-ratios which were considered sufficiently reliable for presentation in court. The number of case reports submitted to the courts per year was 30 in 2005, 59 in 2006, 74 in 2007, and by 2008 it had grown to 98 (personal communication from José Juan Lucena-Molina, February 2009). A commercial version of the IdentiVox system, Batvox, is marketed to other law-enforcement agencies by a spin-off company, Agnitio, with customers in several countries including Chile, China, Colombia, France, Finland, Germany, Malaysia, Mexico, South Korea, and the United Kingdom.

2.4.2 Australia

In Australia forensic-voice-comparison casework is typically conducted by university-based researchers. To date only two likelihood-ratio forensic voice comparison reports have been presented in court, both were acoustic-phonetic and were presented by Philip Rose, one in Victoria in 2007 and one in New South Wales in 2008. At the time of submission of the present paper, in May 2009, members of the Canberra, Sydney, and Madrid research groups, in collaboration with the Australian National Institute of Forensic Science, the Australian Federal Police, Queensland Police, Vitoria Police, Western Australian Police, Guardia Civil, and the Australasian Speech Science and Technology Association, are preparing an application for funding to conduct research and develop the infrastructure necessary to make demonstrably reliable likelihood-ratio forensic voice comparison a practical every-day reality in Australia, similar to the situation in Spain. If funding is granted, the project, lead by Geoffrey Stewart Morrison, will combine acoustic-phonetic and automatic approaches and will include the collection of a database of recordings of 1000+ speakers from different parts of Australia.

2.4.3 Other countries

I have not been able to obtain concrete information on the use of likelihood-ratio forensic voice comparison in casework in other countries. I would be very happy to receive any relevant information on this topic.

2.5 Resistance to the adoption of the likelihood-ratio framework

Saks & Koehler's [1] application of the metaphor of a *paradigm shift* to the current situation in

forensic science, which I have adopted here, can only be a partial metaphor in that not all aspects of Kuhn's [54] description of scientific paradigm shifts are applicable. Saks & Koehler describe it "as a metaphor highlighting the transformation involved in moving from a prescience to an empirically grounded science" (p. 892). In Kuhnian terms the current situation in forensic science might be better described as moving from a *pre-paradigm* to a *normal-science* situation. Buckleton [7] reports difficulty in summarising what he calls the *frequentist approach* since it's definition and logic have never been made explicit by its proponents. While the frequentist approach may appear to be the most promising candidate for a pre-existing paradigm, it is not clear that it ever constituted a single coherent framework accepted as the working paradigm by the majority of forensic scientists.

There are, however, parallels between Kuhnian paradigm shifts and the current situation in forensic science which presumably lead Saks & Koehler [1] to adopt the metaphor. I turn now to one parallel which I find quite striking. According to Kuhn [54], a paradigm shift is typically not completed by the proponents of the new paradigm presenting arguments and empirical evidence which convince all the adherents of the old paradigm. Rather, a paradigm shift is typically completed when its remaining opponents die. Resistance to change is a perfectly understandable aspect of human nature, especially if one has built one's reputation on years of experience working in the old paradigm, or if one has a commercial interest in the continuation of the old paradigm.

Buckleton [7] summarises a number of objections to the adoption of the likelihood-ratio framework in forensic DNA analysis, and makes the case that many of these are based on a lack of understanding of the likelihood-ratio framework, or are problems which equally affect all frameworks. He also makes the case that real difficulties in implementation are not unsurmountable, and in some situations only the likelihood-ratio framework is logically defensible (he in fact uses the label *logical approach* for what I have called the likelihood-ratio framework). I would argue that, because of the differences between DNA and voice data described above (sections 1.5 and 2.1), for forensic voice comparison the likelihood-ratio framework is the only logically defensible framework in all situations.

A lack of understanding of the likelihood-ratio framework also appears to be a factor in the resistance to its adoption for forensic voice comparison. For example, Coulthard & Johnson [55] present a rather negative portrayal of the likelihood-ratio framework, but in the 3.5 pages which they devote to the topic there are 6 inaccuracies. Morrison [56] argues that with a proper understanding of the likelihood-ratio framework, the majority of Coulthard & Johnson's objections can be dismissed.

In what I interpret as in part an attempt to resist pressure to adopt the likelihood-ratio framework while addressing recognised problems in previous and contemporary practice, between 2005 and 2007 a number of forensic speech scientists based in the United Kingdom collaborated on the production of a position statement as to what they considered a correct framework for the evaluation and presentation of forensic-voice-comparison evidence [15]. Although they presented their framework as correctly providing the probability of evidence given competing hypotheses, the framework was inconsistent and in two instances advocated giving categorical posterior-probability statements of exclusion or identification. The framework was a two-stage one, sequentially assessing similarity and typicality, not unlike a framework which had been in use for forensic comparison of DNA profiles prior to it being supplanted by the likelihood-ratio framework [3] (see also Evett [57]). It is also not clear that the framework proposed by the UK-based speech scientists is sufficiently rigorous that it could be empirically tested. For a full critique of the UK framework see Rose &

Morrison [58].

The only principled objections raised by opponents of the adoption of the likelihood-ratio framework for forensic voice comparison appear to be related to defining the relevant population to sample in order to calculate the typicality component of the likelihood ratio, and the cost of the work involved in collecting samples from the relevant population. These were also problems in the development of forensic comparison of DNA profiles, but substantial investment in research and in the development of databases of DNA profiles means that these are now seldom a practical impediment [3]. I see no reasons why, with sufficient investment in research and infrastructure, it should not also be possible to solve these problems with respect to the practical implementation of forensic voice comparison.

3 CONCLUSION

The ongoing paradigm shift in the forensic comparison sciences is a shift towards requiring that evidence be evaluated and presented in a logically correct manner and that the reliability of the results be demonstrable. It is essential to meet these requirements in order to minimise the probability of forensic science contributing to miscarriages of justice. The framework adopted for DNA profile comparison in the 1990s which allows it to meet these requirements is the likelihood-ratio framework. Although forensic voice comparison lags behind forensic DNA analysis in the adoption of the likelihood-ratio framework, it may have a leading position amongst other branches of forensic science. That said, substantial resistance to the adoption of the likelihood-ratio framework remains. It is hoped that recent calls for reform in forensic science in the US National Research Council report to Congress (February 2009) [2] and the Law Commission of England and Wales Consultation Paper (April 2009) [59] will lead to a more rapid adoption of the likelihood-ratio framework, and to additional investment in research and infrastructure which will make demonstrably reliable likelihood-ratio forensic comparison a practical reality in more branches of forensic science in more parts of the world.

ACKNOWLEDGMENTS

The writing of this paper was financially supported by Australian Research Council Discovery Grant No. DP0774115. Thanks to Didier Meuwley, Philp Rose, and Yuko Kinoshita for comments on an earlier draft of this paper.

REFERENCES

[1] M. J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science. Science 309 (2005) 892–895.

[2] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, National Academies Press, Washington, DC, 2009.

[3] L.A. Foreman, C. Champod, I.W. Evett, J.A. Lambert, S. Pope, Interpreting DNA evidence: A review, International Statistics Journal 71 (2003) 473–473.

[4] I.W. Evett, The theory of interpreting scientific transfer evidence. Forensic Science Progress 4 (1990) 141–179.

[5] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Forensic Evidence for Forensic Scientist, 2nd ed, Wiley, Chichester, UK, 2004.

[6] D. J. Balding, Weight-of-evidence for Forensic DNA Profiles, Wiley, Chichester, UK, 2005.

[7] J. Buckleton, A framework for interpreting evidence, in J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC, Boca Raton, FL, 2005, pp. 27–63.
[8] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication 31 (2000) 193-203.
[9] J. González-Rodríguez, A. Drygajlo, D. Ramos-Castro, M. García-Gomar, J. Ortega-García, Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition, Computer Speech and Language 20 (2006) 331–355.
[10] J. González-Rodríguez, P. Rose, D. Ramos, D. Torre, J. Ortega-García, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Transactions on Audio, Speech, and Language Processing, 15 (2007) 2104–2115.
[11] D. Lucy, Introduction to Statistics for Forensic Scientists, Wiley, Chichester, UK, 2005.
[12] B. Robertson, G.A. Vignaux, Interpreting Evidence, Wiley, Chichester, UK, 1995
[13] P. Rose, Forensic Speaker Identification, Taylor and Francis London, UK, 2002.
[14] P. Rose,. Technical forensic speaker recognition, Computer Speech and Language, 20 (2006) 159–191.
[15] J.P. French, P. Harrison, Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, International Journal of Speech, Language and the Law 14 (2007) 137–144.
[16] D. Meuwly, Forensic individualisation from biometric data, Science & Justice 38 (2006) 198–202.
[17] I.W. Evett, G. Jackson, J.A. Lambert, S. McCrossan, The impact of the principles of evidence interpretation on the structure and content of statements, Science & Justice 40 (2000) 233–239.
[18] I.W. Evett, Towards a uniform framework for reporting opinions in forensic science case-work, Science & Justice 38 (1998) 198–202.
[19] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Computer Speech and Language 20 (2006) 230–275.
[20] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravieer, I. Margrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, D.A. Reynolds, A tutorial on text-independent speaker verification, EURASIP Journal on Applied Signal Processing, 4 (2004) 430–451.
[21] D. Ramos Castro, Forensic evaluation of the evidence using automatic speaker recognition systems, PhD dissertation, Universidad Autónoma de Madrid, Madrid, Spain, 2007.
[22] J.S. Gruber, F. Poza, Voicegram Identification Evidence, vol. 54, American Jurisprudence Trials, Westlaw, 1995.
[23] L.M. Solan, P.M. Tiersma, Hearing voices: Speaker identification in court, Hastings Law Journal 54 (2003) 373–435.
[24] S.R. Lewis, Philosophy of speaker identification. Police applications of speech and tape recording analysis, Proceeding of the Institute of Acoustics 6 (1984) 69–77.
[25] A.P.A. Broeders, The role of automatic speaker recognition techniques in forensic investigations, Proceedings of the International Congress of Phonetic Sciences, Stockholm, vol. 3, 1995, pp. 154–161.
[26] P. Rose, Identifying criminals by their voice: The emerging applied discipline of forensic phonetics, Australian Language Matters 5.2 (1997) 6–7.
[27] P. Rose, Differences and distinguishability in the acoustic characteristics of hello in voices of similar-sounding speakers: a forensic–phonetic investigation, Australian Review of Applied

Linguistics 22 (1999) 1–42.
[28] P. Rose, F. Clermont, A comparison of two acoustic methods for forensic speaker discrimination, Acoustics Australia 29 (2001) 31–35.
[29] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Proceedings of RLA2C Workshop: Speaker Recognition and its Commercial and Forensic Applications, 1998, pp. 125–135.
[30] Q.Y. Kwan, Inference of Identity of Source, PhD dissertation, University of California, Berkeley, USA, 1977.
[31] I.W. Evett, J.S. Buckleton, Statistical analysis of STR data, in A. Carraredo, B. Brinkmann, W. Bär (Eds.), Advances in Forensic Haemogenetics, vol. 6, Springer-Verlag, Heidelberg, Germany, 1996, pp. 79–86.
[32] D. Meuwly, A. Drygajlo, Reconnaissance automatique de locuteurs en sciences forensiques: Modélisation de la variabilité intralocuteur et interlocuteur, in Proceedings of 5eme Congres Français d'Acoustique, 2000, pp. 522–525.
[33] F. Nolan, Speaker identification evidence: its forms, limitations and roles, Proceedings of the International Conference on Law and Language: Prospect and Retrospect, 12–15 December 2001, University of Lapland, Levi, Finland, 2001.
[34] A.P.A. Broeders, Forensic speech and audio analysis forensic linguistics: 1998 to 2001 A review, 13th Interpol Forensic Science Symposium, Interpol, Lyon, France, 2001, pp. D2-53–D2-54.
[35] A.P.A. Broeders, Forensic speech and audio analysis forensic linguistics: A review: 2001 to 2004, 14th Interpol Forensic Science Symposium, Interpol, Lyon, France, 2004, pp. 171–188.
[36] D. Meuwly, M. El-Maliki, A. Drygajlo, Forensic speaker recognition using Gaussian mixture models and a Bayesian framework. Proceedings of the COST-250 Workshop, Ankara, Turkey, 1998.
[37] D. Meuwly, A. Drygajlo, Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling, Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, International Speech Communication Association, 2001.
[38] J. González-Rodríguez, J. Ortega-García, J.J. Lucena-Molina, On the application of the Baysian Framework to real forensic conditions with GMM-based systems. Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, International Speech Communication Association, 2001.
[39] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. PhD dissertation, University of Lausanne, Lausanne, Switzerland, 2001.
[40] D.A. van Leeuwen, J.S. Bouten, Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation, Proceedings of Odyssey04: The Speaker and Language Recognition Workshop, Toledo, Spain, International Speech Communication Association, 2004.
[41] F. Botti, A. Alexander, A. Drygajlo, On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition, Forensic Science International 146S (2004) S101–S106.
[42] A. Alexander, D. Dessimoz, F. Botti, A. Drygajlo, Aural and automatic forensic speaker recognition in mismatched conditions, International Journal of Speech, Language and the Law, 12 (2005) 214–234.
[43] A. Drygajlo, Forensic automatic speaker recognition. IEEE Signal Processing Magazine, (2007, March) 132–135.
[44] Y. Kinoshita, Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio

Based Approach Using Formants, PhD dissertation, Australian National University, Canberra, Australia (2001).
[45] P. Rose, The technical comparison of forensic voice samples, in I. Freckelton, H. Selby, (Eds.), Expert Evidence, Thomson Lawbook Company, Sydney, Australia, 2003, ch. 99.
[46] P. Rose, Forensic speaker recognition at the beginning of the twenty-first century: An over-view and a demonstration, Australian Journal of Forensic Sciences, 37.2 (2005) 49-71.
[47] P. Rose, T. Osanai, Y. Kinoshita, Strength of forensic speaker identification evidence: Multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold, Forensic Linguistics 10 (2003) 179–202.
[48] Y. Kinoshita, Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0, International Journal of Speech, Language and the Law 12 (2005) 235–254.
[49] G.S. Morrison, Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/, International Journal of Speech, Language and the Law 15 (2008) 247–264.
[50] G.S. Morrison, Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs, Journal of the Acoustical Society of America 125 (2009) 2387– 2397.
[51] Jessen, M. Forensic phonetics. Language and Linguistics Compass 2 (2008) 671–711.
[52] C. Zhang, G.S. Morrison, P. Rose, Forensic speaker recognition in Chinese: A multivariate likelihood ratio discrimination on /i/ and /y/, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association, 2008, pp. 1937–1940.
[53] C. Zhang, P. Rose, Strength evaluation of forensic speaker recognition evidence based on likelihood ratio approach [in Chinese]. Zheng ju ke xue [Evidence Science], 16 (2008) 337-342.
[54] T.S. Kuhn, The Structure of Scientific Revolutions, University of Chicago Press Chicago, IL, 1962.
[55] M. Coulthard, A. Johnson, An Introduction to Forensic Linguistics: Language in Evidence, Routledge, London, UK 2007.
[56] G.S. Morrison, Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework, in press. [prepublication version available at http://foresnic-voice-comparison.net]
[57] I.W. Evett, Interpretation: A personal odyssey, in C.G.G. Aitken, D.A. Stoney (Eds.), The Use of Statistics in Forensic Science, Ellis Horwood, Chichester, UK, 1991, pp. 9–22.
[58] P. Rose, G.S. Morrison, A response to the UK position statement on forensic speaker comparison, in press. [prepublication version available at http://foresnic-voice-comparison.net]
[59] Law Commission, The Admissibility of Expert Evidence in Criminal Proceedings in England and Wales: A New Approach to the Determination of Evidentiary Reliability, Law Commission, London, UK, 2009. [Available: http://www.lawcom.gov.uk/expert_evidence.htm]