Science and Justice 49 (2009) 298–308

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Forensic voice comparison and the paradigm shift ☆

Geoffrey Stewart Morrison [a,b]

[a] School of Language Studies, Australian National University, Canberra, ACT 0200, Australia
[b] School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 10 May 2009
Received in revised form 13 September 2009
Accepted 23 September 2009

Keywords:
Forensic voice comparison
Forensic speaker recognition
Forensic speaker identification
Paradigm shift
Likelihood ratio
Reliability

ABSTRACT

We are in the midst of a paradigm shift in the forensic comparison sciences. The new paradigm can be characterised as quantitative data-based implementation of the likelihood-ratio framework with quantitative evaluation of the reliability of results. The new paradigm was widely adopted for DNA profile comparison in the 1990s, and is gradually spreading to other branches of forensic science, including forensic voice comparison. The present paper first describes the new paradigm, then describes the history of its adoption for forensic voice comparison over approximately the last decade. The paradigm shift is incomplete and those working in the new paradigm still represent a minority within the forensic–voice-comparison community.

1. The new paradigm in forensic science

1.1. A paradigm shift

Today we are in the midst of what Saks and 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 objects of known and questioned origin, e.g., DNA profiles, finger marks, hairs, fibres, glass fragments, tool marks, handwriting, and voice recordings. Saks and Koehler point out that they "use the notion of paradigm shift not as a literal application of Thomas Kuhn's concept [2], but as a metaphor highlighting the transformation involved in moving from a pre-science to an empirically grounded science" (p. 892). In Kuhnian terms, Saks and Koehler's paradigm shift might be better described as a shift from a pre-paradigm period towards a period where there is for the first time a single unifying paradigm for conducting normal science, i.e., a shift from a period during which a number of different schools pursue solutions to different sets of problems (with only partial overlap between sets) using different incompatible frameworks, towards a period during which there is agreement throughout the scientific community as to which problems are important (often a superset of the problems addressed by two or more of the pre-paradigm schools), and agreement as to the general procedures for solving these problems and the nature of suitable solutions. Whereas during the pre-paradigm period scientists must address a general audience and explain their theories from the beginning, during a normal-science period scientist principally address an audience which has already been educated in the fundamentals of the paradigm (e.g., by completing at least a bachelor of science degree) and they can immediately focus their efforts on a particular small question which forms part of the larger puzzle. Research efficiency and productivity is therefore greater during a normal-science period than during a pre-paradigm period.

Kuhn uses the term "paradigm" in two different senses, one broader and the other narrower: "On the one hand, it stands for the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community. On the other, it denotes one sort of element in that constellation, the concrete puzzle–solutions which, employed as models or examples, can replace explicit rules as the basis for the solution of the remaining puzzles of normal science." [3] (p. 175). I will essentially be using the broader sense of "paradigm", which subsumes its narrower sense. Although I believe that Kuhn's thinking on scientific revolutions provides a useful tool for understanding the current situation in forensic science, and I point out a number of parallels below, one does not find a 100% correlation. One reason for this may be that forensic science is an applied science which must serve the imminent needs of society, and this consideration impinges to a greater extent than is the case in the natural sciences. In this, the forensic scientist is more like an engineer: "Unlike the engineer, and many doctors, and most theologians, the scientist need not choose problems because they urgently need solution and without regard for the tools available to solve them." [2] (p. 163).

Saks and Koehler [1] propose that a paradigm shift has already occurred in DNA profile comparison, and that other forensic comparison sciences are now shifting towards the new paradigm. In the present

☆ This is a revised version of an invited presentation given at the 2nd International Conference on Evidence Law and Forensic Science, Beijing, China, 25–26 July, 2009.

E-mail address: geoff.morrison@anu.edu.au.

1355-0306/$ – see front matter
doi:10.1016/j.scijus.2009.09.002

paper my aim is to first describe the characteristics of the new paradigm, and then tell the story of its adoption to date in the field of forensic voice comparison.

1.2. The new paradigm

Saks and Koehler [1] describe the new paradigm as "empirically grounded science" (p. 892) as exemplified by "data-based, probabilistic assessment" (p. 893) as is current practice in forensic DNA comparison. They recommend that other forensic comparison sciences emulate DNA comparison, including that they "construct databases of sample characteristics and use these databases to support a probabilistic approach" (p. 893). They also make it clear that another important aspect of the new paradigm is the quantification and reporting of the limitations of forensic comparison via the measurement of error rates. The new paradigm therefore echoes the requirements for admissibility of scientific evidence set out in the US Supreme Court ruling in *Daubert v Merrell Dow Pharmaceuticals* (92–102) 509 US 579 [1993], which Saks and Koehler identify as a driving force for the paradigm shift. The Court ruled that, when considering the admissibility of scientific evidence, the judge must consider the methodology's scientific validity and evidentiary reliability, including whether it has been empirically tested and found to have an acceptable error rate. The call for other branches of forensic science to be more "scientific", emulate DNA profile comparison, and conform to the *Daubert* requirements was recently reiterated in the February 2009 release of the National Research Council (NRC) report to Congress on Strengthening Forensic Science in the United States [4]. Important aspects of a scientific approach identified in the report include "the careful and precise characterization of the scientific procedure, so that others can replicate and validate it; . . . the quantification of measurements . . .; the reporting of a measurement with an interval that has a high probability of containing the true value; . . . [and] the conducting of validation studies of the performance of a forensic procedure" (p. 4-8); the latter requiring the use of "quantifiable measures of the reliability and accuracy of forensic analyses" (p. S-16). The NRC report clearly recommends the use of more objective analytic methodologies over more subjective experience-based methodologies.

Although there does not appear to be any indication that either set of authors were consciously aware of this, there is one other component of the new paradigm which I believe is implicit in Saks and Koehler's [1] and the NRC report's [4] recommendation that other forensic comparison sciences emulate forensic DNA comparison: the adoption of the *likelihood-ratio framework* for the evaluation of evidence. In fact the NRC report consistently describes "identification" and "individualisation" as the aims of forensic science, which is antithetical to the use of the likelihood-ratio framework (see Section 1.4 below). The term "likelihood ratio" appears only once in the report, and this is in the title of a cited paper; however, the report recommends Aitken and Taroni [5], Evett [6], and Evett et al. [7] as providing "the essential building blocks for the proper assessment and communication of forensic findings"(p. 6-3), and all three advocate the use of the likelihood-ratio framework.

1.3. The likelihood-ratio framework

The leading *rôle* of forensic DNA comparison in the paradigm shift can in large part be attributed to the fact that it is a relatively new branch of forensic science which was put under extensive scrutiny when it was first presented in court in the late 1980s and early 1990s, and to the fact that it was developed by researchers who were trained and experienced in modern approaches to scientific research. The strong modern scientific background of those working in forensic DNA analysis arguably made it easier for them to understand and ultimately adopt what many forensic statisticians recommend as the logically correct framework for the evaluation of comparison evidence, the *likelihood-ratio framework*. Descriptions of the likelihood-ratio framework can be found in numerous textbooks and articles including Aitken and Taroni [5], Balding [8], Buckleton [9], Evett [10], Lucy [11], Robertson and Vignaux [12], and with specific reference to forensic voice comparison Champod and Meuwly [13], González-Rodríguez et al. [14], González-Rodríguez et al. [15], and Rose [16,17]. For a history of developments in forensic statistics prior to the advent of forensic DNA analysis (including use of the likelihood-ratio framework) see Evett [6], and for a history of statistical procedures applied to the evaluation of DNA evidence and the ultimate adoption of the likelihood-ratio framework in that field see Foreman et al. [18].

What follows is a brief description of the likelihood ratio framework. For simplicity, the description is provided only at the source level as this is the most relevant level for forensic voice comparison (see Cook et al. [19] on the hierarchy of *source*, *activity*, and *offence* propositions). The activity level is seldom important in forensic voice comparison because issues of transfer and persistence are seldom pertinent: voice recordings are usually deliberately recorded, and those presented for forensic analysis are typically associated with warrants and chain-of-custody documentation. Authentication of audio recordings, and analysis of disputed utterances, are normally considered to be areas of expertise which are distinct from forensic voice comparison. In forensic voice comparison one must, however, consider the effects of the conversion of the acoustic signal to an electronic signal and often its transmission over a telephone system, which can result in relatively poor quality voice recordings and potentially mismatches between the recording quality of known and questioned samples (transmission-channel effects). There may also be differences in speaking style, e.g., a lively telephone conversation on the recording of the questioned voice, and subdued answers to questions asked in a police interview on the recording of the known voice. The outcome of a forensic voice comparison may be of direct relevance for the offence propositions, for example, if the offence is uttering death threats and the questioned voice recording is a recording of someone uttering death threats.

In the likelihood-ratio framework the task of the forensic scientist is to provide the court with a *strength-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 Eq. (1).

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

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 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 strength of evidence, the forensic scientist must consider not only the degree of similarity between the samples, but also their degree of typicality 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 in terms of some physical properties, 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 lead to relatively greater support for the same-origin hypothesis, and less similarity and more typicality lead to relatively greater support for the different-origin hypothesis.

The likelihood-ratio framework is a conceptual framework which can be applied to subjective experience-based beliefs as to the likelihoods of the evidence given the competing hypotheses; however, to implement the data-based and quantitative-measurement aspects of the new paradigm, the forensic scientist must have access to a database of samples which are representative of the relevant population in order to calculate a quantitative estimate of the typicality of the known and questioned samples. The relevant population is the population to which the offender belongs. Practically, this is probably less than the entire population of the planet, it could be confined to a particular geographical area, a particular ethnic group, or, in forensic voice comparison, speakers of a particular language and dialect. Selection of an appropriate population to sample is not a simple matter, see discussion in [5] (pp. 274–271) and [11] (pp. 129–133).

1.4. 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 strength-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 strength/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 fingermark samples, etc., and the trier of fact will weigh all of these together. Not all the evidence will be forensic comparison evidence presented as likelihood ratios, and the trier of fact must also consider the weight of other evidence such as eye-witness testimony. In addition, before any evidence has been presented the trier of fact will have some belief as to the innocence/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 Eq. (2).

$$\underbrace{\frac{p(H_{so}|E)}{p(H_{do}|E)}}_{\text{Posterior odds}} = \underbrace{\frac{p(E|H_{so})}{p(E|H_{do})}}_{\text{likelihood ratio}} \times \underbrace{\frac{P(H_{so})}{P(H_{do})}}_{\text{prior odds}}. \tag{2}$$

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, when conducting their analysis, 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, then pragmatic prior odds could be 1/99; 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 not-guilty would still be 1/99 immediately prior to the presentation of the likelihood ratio from the forensic evidence in question).

It is inappropriate for the forensic scientist to present the posterior odds because the posterior odds include information and assumptions from sources other than a scientific evaluation of the known and questioned samples. If the forensic scientist were to present posterior odds then they would have to supply their own prior odds, and 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. Human bias was a major concern in the NRC report [4] (pp. 4-9–4-11). It is a strength of the likelihood-ratio framework that it is resistant to influence from such sources of bias.

Although the likelihood ratio is a component of Bayesian analysis, I have used the term "likelihood-ratio framework" rather than "Bayesian framework" since the latter, unlike the former, could imply that the forensic scientist makes use of priors and calculates posteriors [9,13,17]. An alternative to "likelihood-ratio framework" used by some authors (e.g., [9]) is "logical approach", I prefer "likelihood-ratio framework" because I believe it is more transparent. It should also be noted that the fact that forensic scientists present likelihood ratios in court does not imply that the trier of fact must assign numeric weights to evidence which is not forensic comparison evidence nor that they must arrive at their decision via the rigid application of a formula such as Eq. (2) (*R v Adams* [1996] EWCA Crim 222, *R v Adams* [1997] EWCA Crim 2474, *R v GK* [2001] NSWCCA 413, [8] (pp. 149–151), [9,20,21]).

A terminological point which arises from the discussion above is that in the likelihood-ratio framework the forensic scientist does not perform "identification" or "individualisation", because these terms imply determining a posterior probability (see Meuwly [22] on terminological and logical problems with the use of the terms "identification" and "individualisation" in forensic science). A neutral term such as "comparison" is more appropriate [23]. I therefore use the term "forensic voice comparison" rather than either of the traditional terms "forensic speaker identification" and "forensic speaker

recognition" ("recognition" also implies a posterior probability, note also that "speaker comparison" would be akin to calling fingermark comparison "toucher comparison"). Following Meuwly's logic I should actually be using a term such as "forensic comparison of voice recordings", i.e., it is the properties of the recordings which are actually compared, not the voices themselves. 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. Measuring reliability

The reliability of the output of a forensic comparison system can be assessed by testing it on a large number of pairs of samples where it is known for each pair whether its members have the same origin or different origins, then comparing the system's output with this knowledge about the input. Saks and Koehler [1] and the NRC report [4] (pp. 4-5–4-9) describe quantitative reliability in terms of identification error rates, i.e., false positives (different-origin pairs declared to be same origin) and false negatives (same-origin pairs declared to be different origin). Identifications are based on posterior probabilities and this approach is therefore inconsistent with the likelihood-ratio framework. 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 strength 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. It is sometimes convenient to convert likelihood ratios to log likelihood ratios since the latter are symmetrical about zero, e.g., likelihood ratios of 1000 (1000 in favour of the same-origin hypothesis) and 1/1000 (1000 in favour of the different-origin hypothesis) become log-base-ten likelihood ratios of +3 and −3 respectively. Ideally, for a same-origin pair the forensic comparison system should produce a large positive log likelihood ratio, and for a different-origin pair it should produce a large negative log likelihood ratio. For a same-origin comparison, a small positive log likelihood ratio is not as good as a large positive log likelihood ratio, a small negative log likelihood ratio is worse than a small positive log likelihood ratio, and a large negative log likelihood ratio is worse than a small negative log likelihood ratio (*mutatis mutandis* for a different-origin comparison). Small and large negative likelihood ratios would respectively provide weak and strong support for the different-origin hypothesis when it is known that in fact a same-origin pair was being tested. It is worse to report a likelihood ratio of 1000 in favour of a contrary-to-fact hypothesis than it is to report a likelihood ratio of 10 in favour of a contrary-to-fact hypothesis because the former has greater potential to contribute towards a miscarriage of justice.

The log-likelihood-ratio cost (C_{llr}) [24–26] is a measure of the reliability of a system which outputs likelihood ratios. C_{llr} was developed for use in automatic speaker recognition and has subsequently been applied to forensic voice comparison [15,27–29]. To calculate C_{llr}, one must first calculate a C_{llr} component value for the likelihood ratio from each test pair. Fig. 1 provides a plot of the function for calculating a C_{llr} component value when the input to the system is a same-origin pair. Large positive log-likelihood-ratio values which correctly support the same-origin hypothesis are assigned very low C_{llr} component values, log-likelihood-ratio values close to zero provide little support for either the same-origin or different-origin hypothesis and are assigned moderate C_{llr} component values, and negative log-likelihood-ratio which contrary-to-fact support the different-origin hypothesis are assigned high C_{llr} component values which increase rapidly as the log-likelihood-ratio values become more negative and provide stronger contrary-to-fact support for the different-origin hypothesis. The function for calculating a C_{llr} component value when the input to the system is a different-origin pair is a mirrored version of Fig. 1, mirrored at log likelihood ratio = 0. To calculate C_{llr}, one finds the mean of all the C_{llr} component values from same-origin pairs, the mean of all the C_{llr} component values from different-origin pairs, and then takes the mean of the latter two means. C_{llr} therefore provides a quantitative measure of the general reliability of a forensic comparison system. 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. It is important to note that (as with other measures of reliability such as identification error rates) C_{llr} depends on the test data as well as the forensic comparison system, hence it serves better as a relative measure of reliability rather than an absolute measure. Also, to be meaningful in casework, the quantity and quality of each test pair should be matched as closely as possible to the quantity and quality of the known and questioned samples, e.g., for voice recordings this would include attempting to match duration, recording quality, and speaking style.

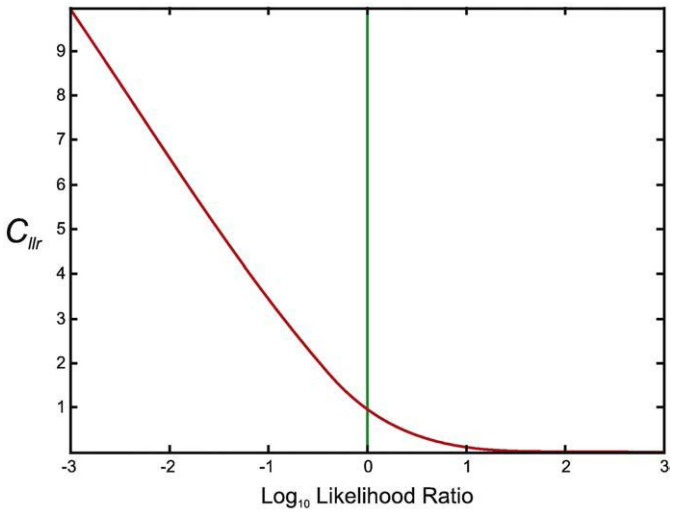

Fig. 1. Plot of the function for calculating a C_{llr} component value for a same-origin comparison.

Within the likelihood ratio framework it is also possible to report an error rate for the specific likelihood ratio which is calculated for the comparison of the known and questioned samples. For example, 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 pairs in the test data which resulted in likelihood ratios of equal to or greater than 100.

An additional issue related to reliability in the new paradigm is the "reporting of a measurement with an interval that has a high probability of containing the true value" [4] (p. 4-8). Although nothing on this issue with respect to forensic voice comparison has yet been published, some preliminary research is underway on the calculation of credible intervals for likelihood ratios.

2. Forensic voice comparison and its place in the paradigm shift

2.1. 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*. Of these it is the latter two which are most appropriate for use in the new paradigm. For simplicity of exposition the four approaches 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 description of each approach below is meant to be a rough sketch rather than a thorough exposition; fuller descriptions are provided in Jessen [30] and Rose [16], and for more detail on automatic approaches see Bimbot et al. [31] and Ramos Castro [29].

2.1.1. 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. Audible features which are exploited could be the sorts of differences which distinguish dialects, e.g., consider the way the word "height" (phonemically transcribed /haɪt/) would be pronounced by English speakers from the US Mid-West, Southern US, Canada, and Australia (in broad phonetic transcription these could be [haɪt], [haːt], [hʌɪt], and [hɐɪt] respectively). Such large dialectal differences are often salient even to the untrained listener, but an expert trained in auditory phonetics will be able to notice and systematically label smaller idiolectal differences. Audible features could also be related to vocal-fold activity, e.g., whether the voice is breathy (like Marilyn Munroe) or creaky (like Louis Armstrong), or could be what might be considered speech impediments of varying severity, e.g., pronouncing "r"s as "w"s (/r/ as [w]). See Jessen [30] and Rose [16] for more examples. Although there may be some auditory features whose frequency of occurrence can be quantified and data-based likelihood ratios calculated (see discussion in [32]), in general the auditory approach is experience-based and not consistent with the new paradigm. Although theoretically it would be possible to obtain a measure of the reliability of a practitioner of the auditory approach by having them compare a large number of pairs of samples known by the tester (but not the testee) to be of same or different origin, as far as I am aware no large-scale tests of a pure auditory approach have been conducted.

2.1.2. 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 in a format known as a *spectrogram*. 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 Fig. 2. The forensic use of spectrograms was first described in press in Kersta [33] in 1962. There was much debate about the validity of the spectrographic approach during the 1960s, 70s, and 80s. Although it may still have some diehard supporters, the general conclusion of the scientific community is that the spectrographic approach is not scientific and not reliable. 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 . 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). For summaries of the historical debate about the validity of the spectrographic approach see Gruber and Poza [34], Rose [16] (pp. 107–122), and, from a legal perspective, Solan and Tiersma [35]. Also see Schwartz [36] on why voiceprinting won't go away. From the perspective of the new paradigm, it is worth noting that an important component of the debate about the validity of the spectrographic approach was its reliability as measured by identification error rates, and large-scale testing was conducted.

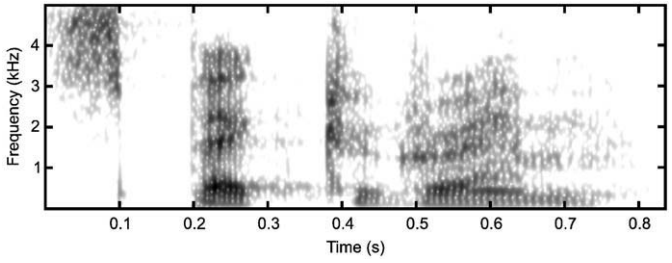

Fig. 2. Example of a spectrogram.

2.1.3. Acoustic–phonetic approach

The acoustic–phonetic approach was developed by phoneticians trained in acoustic phonetics and involves making quantitative measurements of the acoustic properties of speech sounds. Typically, comparable phonetic units are identified in both known and questioned voice samples and then the 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 acoustic properties of many of the features used in the auditory approach can also be quantitatively measured to provide acoustic–phonetic features. Some acoustic–phonetic features, such as fundamental frequency (the acoustic correlate of pitch) and the second formant, have the advantage that they are relatively robust to transmission–channel effects. Acoustic measurements are made using signal-processing algorithms but with substantial human-expert supervision aimed at obtaining high measurement accuracy. The time and expense involved in data analysis is a major constraint on the application of the acoustic–phonetic approach.

2.1.4. Automatic approach

The automatic approach was developed by signal-processing engineers. As with the acoustic–phonetic approach, it is based on quantitative measurements 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 attempted here). Typical automatic features are not particularly robust to transmission–channel effects, but substantial research has been conducted on statistical procedures to compensate for transmission–channel mismatch. 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.

2.1.5. Relative popularity of different approaches

Between 2004 and 2005 Tina Cambier-Langevald of the Netherlands Forensic Institute (NFI) conducted an exercise in which 12 participants submitted their analyses of the same voice samples and described their approach [37]. Although this may not constitute a large random sample, it provides some idea of the relative popularity of each approach among researchers and practitioners. Cambier-

Langevald's classification was somewhat different to mine, but, as far as I am able to ascertain, 5 participants used what I would describe as auditory–acoustic–phonetic approaches, 4 used what I would describe as acoustic–phonetic approaches, 2 used automatic approaches, and 1 used a spectrographic approach. Even within the different approaches there was extensive heterogeneity in the choice of what parts of the recordings to measure, the features measured, measurement and analysis techniques, and the evaluation and reporting of results. Only 4 of the 12 participants reported likelihood ratios (both the participants who used automatic approaches and two of the participants who used acoustic–phonetic approaches).

2.2. Differences between DNA and voice data

The following discussion includes a simplified account of DNA profile comparison, my purpose is to highlight some basic differences between DNA and voice data, not to discuss issues in the interpretation of DNA evidence (readers interested in the latter may wish to consult resources such as [8] and [38]). DNA profile data consist of discrete values (e.g., counts of short tandem repeats) from a finite number of measurements (e.g., pairs of alleles at specific loci). DNA properties are discrete at the molecular level, their values are continuous at the measurement level (which can, for example, be graphically represented as the location and height of peaks on an electropherogram), but they are typically converted back to discrete values to provide data for statistical analysis. It is the latter to which I refer when I use the term "DNA profile". For simplicity I will assume (unrealistically) that DNA profiles have no measurement errors, that samples are not contaminated, that the organisms from which DNA samples originate have not undergone transplants, etc. It is possible to obtain a "match" between two DNA profiles, i.e., for each corresponding locus and allele each of the two profiles has the same discrete value. Under the assumptions laid out above, the DNA profile of an individual organism does not change from occasion to occasion, hence the probability of obtaining matching DNA profiles given the same-origin hypothesis is 1, and the probability of obtaining non-matching DNA profiles given the same-origin hypothesis is 0. The numerator of the likelihood ratio is therefore either 1 or 0 [5] (p. 404), [10]. If the two samples do not match, the numerator of the likelihood ratio is 0 and the denominator is irrelevant, the value of the likelihood ratio is 0 and via Bayes' Theorem the posterior odds will also be 0, the two samples do not have the same origin. If the two samples match, the numerator of the likelihood ratio is 1 and the size of the likelihood ratio is then dependant on the denominator, the probability of the DNA profile of the questioned sample matching the DNA profile of the known sample if the questioned sample comes from a source other than the known organism. Often the "match probability" rather than the likelihood ratio is reported in court (*R v Doheny & Adams* [1996] EWCA Crim 728 directed DNA experts to provide match probabilities, see also [10] and [8] pp. 151–153), this is simply the denominator of the likelihood ratio, or equivalently the inverse of the likelihood ratio given in Eq. (1), i.e., it is the probability of obtaining the matching DNA profile in question under the different-origin versus the same-origin hypothesis [8] (p. 24), [18] (p. 484).

An acoustic–phonetic or automatic forensic–voice-comparison system would be based on measurements of acoustic properties of voices. These acoustic properties are continuous, not discrete. There is also substantial within-speaker variation, even if a speaker says exactly the same words twice in a row it would be extremely unlikely for there not to be measurable differences in the acoustic properties of the two utterances. Note that this is not just the precision of the measurement techniques, it is also intrinsic variability in the source. In practice a speaker is unlikely to repeat long stretches of exactly the same words, and there will likely also be variability due to factors such as phonetic context and speaking style (and also often transmission–channel). For continuously valued properties with this sort of variation a "match", in terms of two samples being indistinguishable within the precision of measurement techniques, or in terms of not having (at some alpha level) a statistically significant difference for the combination of intrinsic and measurement variability, or in terms of some pre-determined difference threshold (whether experience or empirically based), suffers from a cliff-edge effect [12] (pp. 118–120), [39], [40]. "Match" is therefore not a useful concept for the acoustic properties of voices (the same can probably be said for the properties of objects of comparison in many other branches of forensic science). The numerator of a likelihood ratio calculated from a forensic voice comparison cannot therefore be either 0 or 1, a match probability cannot be calculated, and the results must be reported in the form of a full likelihood ratio. Some would argue that, since the simplifications made above with respect to DNA profile comparison are not valid, DNA results should also be reported as full likelihood ratios (personal communication from Didier Meuwly, April 2009, see [41] §30:41 on problems with the process of converting continuous electropherogram data to discrete values).

2.3. The adoption of the new paradigm by the research community

2.3.1. Proposals to adopt the likelihood-ratio framework

The first published proposal that the likelihood-ratio framework be adopted for forensic voice comparison appears to have been made by S. R. Lewis in 1984 [42]. 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 [43]. 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 [44–46]. Rose cites Robertson and 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 journal article submitted to *Speech Communication* in October 1998 and published in September 2000 [13,47]. 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 [48], Lewis [42], and Evett and Buckleton [49] 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 *Congrès Français d'Acoustique* in September 2000 [50]. 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 quantitative data-based implementation of the framework [51]. 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 [52,53]. 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 new paradigm in automatic forensic voice comparison

The first data-based 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 which calculated likelihood ratios [54]. 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). However, 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 Andrzej 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 [55,56]. Didier Meuwly's PhD dissertation was also completed in 1999 and published in 2001 [57].

Since then, data-based implementation of 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 [58], and (although their goal is not primarily forensic) evaluation via likelihood-ratio based C_{llr} was adopted by the US *National Institute of Standards and Technology Speaker Recognition Evaluations* (NIST SRE) in 2006.

Important journal articles describing the likelihood-ratio framework and its use for the calculation of data-based likelihood ratios in automatic forensic voice comparison were published by the Lausanne and Madrid groups in the middle of the decade [14,15,59–61].

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 new paradigm in acoustic–phonetic forensic voice comparison

In acoustic–phonetic forensic voice comparison, data-based implementation of the likelihood-ratio framework has been pioneered by a research group working in Canberra, Australia. The first such implementation was in Yuko Kinoshita's PhD dissertation completed in 2001 [62]. In 2002 and 2003 Philip Rose published a book and a book chapter on likelihood-ratio forensic voice comparison, the first aimed primarily at phoneticians [16], and the second aimed primarily at the legal community [32]. Although now somewhat dated, Rose [16] 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 [17,63], and journal articles by the Canberra group reporting research results obtained using data-based implementations of the framework include [27,64–66]. Recently Cuiling Zhang of the *China Criminal Police University* in Shenyang has collaborated with the Canberra group, developing the first data-based acoustic–phonetic likelihood-ratio forensic voice comparison of Chinese speech [67,68].

A survey of forensic phonetics authored by Michael Jessen of the *Bundeskriminalamt* (BKA, German Federal Criminal Police Office) was published in 2008. Within this, Jessen recommends the adoption of the likelihood-ratio framework [30]. Unlike the situation in the automatic forensic voice comparison community, in the acoustic–phonetic forensic voice comparison community those working in the new paradigm remain a minority.

2.3.4. Combination of automatic and acoustic–phonetic approaches within the new paradigm

There is increasing interest in combining aspects of the automatic and acoustic–phonetic approaches to forensic voice comparison within the new paradigm. Philip Rose and Geoffrey Stewart Morrison of the Canberra group are currently working on a research project on this topic funded by the *Australian Research Council* from 2007 to 2010. 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's first publication on forensic voice comparison was by Tharmarajah Thiruvaran, Eliathamby Ambikairajah, and Julien Epps [69]). Another project investigating automatic and acoustic–phonetic approaches to forensic voice comparison is a collaboration between the BKA, the Romanian Ministry of Justice, and the Austrian Academy of Science, funded by the European Union from 2008 to 2010 (their first publication from this project was by Timo Becker, Michael Jessen, and Catalin Grigoras [70]). Also, a special session on combining automatic and acoustic–phonetic approaches was organized 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 [15,27]. The judicial phonetics specialisation in the Masters in Phonetics and Phonology programme run by the *Consejo Superior de Investigaciones Científicas* (Spanish National Research Council) and the *Universidad Internacional Menéndez Pelayo* since 2008 now includes training in both acoustic–phonetic and automatic forensic voice comparison within the new paradigm.

2.4. The adoption of the new paradigm by the forensic practitioner, law-enforcement, and judicial communities

2.4.1. Spain

The only jurisdiction where forensic voice comparison can be said to be commonly practised using data-based implementation of 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 the *Guardia Civil* considered sufficiently reliable for submission to court. The number of case reports submitted to the courts per year was 30 in 2005, 59 in 2006, 74 in 2007, and 98 in 2008 (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 data-based 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. In non-judicial writings, The Honourable David Hargraves

Hodgson, Judge of Appeal of the Supreme Court of New South Wales, has commented favourably on the use of Bayesian approaches for the evaluation and presentation of forensic evidence, including forensic–voice-comparison evidence [71,72]. At the time of writing (September 2009) members of the Canberra, Sydney, and Madrid research groups, in collaboration with the Australian National Institute of Forensic Science, and the forensic laboratories of the Australian Federal Police, Victoria Police, Western Australia Police, and other partners are preparing an application for funding to conduct research and develop the necessary infrastructure with the aim of making likelihood-ratio forensic voice comparison of demonstrable reliability a practical every-day reality in Australia. If funding is granted, the project will combine acoustic–phonetic and automatic approaches and will include the collection of a database of recordings of approximately 1000 speakers from different parts of Australia.

2.4.3. Other countries

I have not been able to obtain concrete information on the adoption of the new paradigm for 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 paradigm shift

According to Kuhn [2] (ch. 12), 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 (pp. 150–151). 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. But resistance to change may also be based on a genuine belief that the old paradigm will ultimately lead to solutions for all the outstanding problems and that a paradigm shift is not warranted. Indeed if scientists were too quick to adopt potential new paradigms, the scientific community would be in constant flux and one would not observe long periods of productive normal science.

Given Kuhn's [2] observations (published in 1962), it is not surprising to find that there has been considerable resistance to the paradigm shift in forensic comparison sciences. D. V. Lindley's proposal to implement a full Bayesian framework at the 1977 *Royal Statistical Society / Institute of Statisticians Conference* met some vehement opposition: "I believe Lindley's suggestion is not only mad, it is extremely dangerous" R. A. Carr Hill [73] (p. 216). I. W. Evett [39] reported having had great difficulty getting his initial work on Bayesian approaches accepted in the 1980s: "A paper which I submitted . . . was savaged by the referees and rejected without a single word of encouragement. A paper which I presented at a colloquium . . . met a response which bordered on the abusive. . . [; however,] When, several years later, I did succeed in having a Bayesian paper published . . . it was given the . . . Award for the best paper of the year!" (p. 12). Evett [39] describes his own conversion experience in the 1970s, including discussions with D. V. Lindley in which he initially defended a two-stage frequentist statistical approach to data-based forensic comparison of glass fragments, but ultimately came to be an advocate of Bayesian approaches. He also describes a feeling of *déjà vu*, when exactly the same issues came up again on the advent of forensic DNA profile comparison towards the end of the 1980s. As mentioned above (Section 2.3.2), in the late 1990s the Lausanne forensic–voice-comparison group also experienced hostility from an audience member in response to a conference presentation, and negative reviews to articles submitted to journals.

Buckleton [9] 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 argues that real difficulties in implementation are not unsurmountable, and in some situations only the likelihood-ratio framework is logically defensible. He reports difficulty in summarising what he calls the *frequentist approach* since its 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.

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 and forensic linguistics. For example, Coulthard and Johnson [74] present a rather negative portrayal of the likelihood-ratio framework, particularly critical of Rose's work, but in the 3.5 pages which they devote to the topic there are 6 inaccuracies. Morrison [21] argues that with a proper understanding of the likelihood-ratio framework, the majority of Coulthard and Johnson's objections can be dismissed. Watt [75] writes that "Speaker comparison . . . is based upon close comparison of two speech samples with a view to estimating the likelihood that the samples were produced by the same person" and "any judgment we make about the degree of correspondence between two speech samples not known in advance to have been produced by the same person should, where feasible, be cast in terms of the statistical likelihood that they came from the same speaker (Rose, 2006)." These quotes are examples of the prosecutor's fallacy ([5] pp. 79–82, [8] pp.146–147, [76]), and the mis-citation of Rose [17] demonstrates a lack of understanding of Rose's description of the likelihood-ratio framework.

According to Kuhn [2], a paradigm shift is typically precipitated by a widespread awareness of there being a crisis, an acknowledgment by a large number of scientists that there are serious outstanding problems which it does not appear possible to solve within the existing paradigm. In forensic voice comparison the source of the crisis appears to be largely external, being driven by judicial rulings such as *Daubert*, *Adams*, and *Doheny and Adams*; developments in other branches of forensic science, particularly DNA profile comparison; and reviews, recommendations, and standards such as the NRC report [4], the Law Commission of England and Wales Consultation Paper [77], and the Association of Forensic Science Providers' Standards for the Formulation of Evaluative Forensic Science Expert Opinion [78]. The existence of a crisis was acknowledged by a number of forensic speech scientists based in the United Kingdom, who between 2005 and 2007 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 [23]. They did not, however, adopt the new paradigm which I have described here. Indeed, I interpret their actions as an attempt to resist pressure to adopt the new paradigm and instead create and promote an alternative paradigm which is closer to their previous practice and thus easier for them to implement. Although they present their framework as correctly providing the probability of evidence given competing hypotheses, the framework is inconsistent and in two instances advocates giving posterior-probability statements of exclusion or identification: "we see no logical flaw in making the statement that the samples are spoken by different speakers." (p. 141), "we consider it justified to make categorical statements of identification." (p. 142). The framework is a two-stage one, sequentially assessing similarity and typicality, reminiscent of frameworks which had been in use in other forensic comparison sciences, including DNA, before being supplanted by the likelihood-ratio framework [18,39]. Reliability is not mentioned in the UK position statement, and I have not seen any publications which test the reliability of forensic voice comparison conducted using the UK framework. For a full critique of the UK position statement see Rose and Morrison [40].

The UK position statement [23] ends by saying: "Finally, we accept in principle the desirability of considering the task of speaker comparison in a likelihood ratio (including Bayesian) conceptual framework.

However, we consider the lack of demographic data along with the problems of defining relevant reference populations as grounds for precluding the quantitative application of this type of approach in the present context." (p. 142). Given this, it is unclear why the authors of the UK position statement did not adopt an implementation of the likelihood-ratio framework using experience-based estimates of the probability of the evidence given competing hypotheses. At least one signatory of the UK position statement explicitly rejects this possibility: "Where it is not possible to express an opinion in this way – which is in reality almost always, because in most cases we lack population statistics on the distribution of speech features even in well-described languages like English – the use of likelihood statistics should be avoided altogether." [75]. The likelihood-ratio framework is a conceptual framework, not in and of itself dependant on data, and an experience-based implementation of the likelihood ratio framework would be defensible if it were coupled with reliability testing. "For an expert to say 'I think this is true because I have been doing this job for *x* years' is, in my view, unscientific. On the other hand, for an expert to say 'I think this is true and my judgement has been tested in controlled experiments' is fundamentally scientific." Evett [39] (p. 21). I consider demonstrable reliability to be an essential aspect of the new paradigm, and quantitative data-based analysis a highly desirable aspect; if an experience-based human expert can be demonstrated to produce likelihood ratios of greater reliability than a quantitative data-based system, then I would prefer the experience-based system over the data-based system. Also, although I am a proponent of quantitative data-based implementation of the likelihood-ratio framework, I can envisage exceptional circumstances in which it would be essentially impossible to collect meaningful population data, but in which experienced-based testimony would be of value to the court.

Although the bulk of the UK position statement [23] seems to be concerned with offering an alternative to the likelihood-ratio framework component of the new paradigm, the following quotation is instead a rejection of the data-based component: "we consider the lack of demographic data along with the problems of defining relevant reference populations as grounds for precluding the quantitative application of this type of approach in the present context."(p. 142). This is not just a rejection of data-based implementation of the likelihood-ratio framework, but also a rejection of all data-based frameworks (because of the data-collection problem), and a rejection of all frameworks which consider typicality (because of the defining-relevant-population problem), which logically would include the UK framework itself. The problems of 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 and analysing samples from the relevant population are real problems which must be addressed ([5] pp. 274–271, [11] pp. 129–133). 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 problems are now seldom a practical impediment to casework [18]. 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 in the new paradigm. This will clearly be harder in places like the UK with great dialectal heterogeneity compared to places like Australia with relative dialectal homogeneity, but if one accepts any data-based paradigm then this is an invitation to conduct more research rather than abandon the paradigm. In fact greater dialectal heterogeneity has the potential to ultimately lead to forensic voice comparison being more valuable for the trier of fact: Dialectal heterogeneity could lead to greater between-speaker variation with the potential for larger likelihood ratios, or could lead to a reduction in the size of the potential population of offenders considered by the trier of fact.

Finally, it is my experience that some opponents of the adoption of the new paradigm are under the mistaken belief that its proponents claim that it is an immediate solution to all existing problems, and the opponents therefore dismiss the new paradigm because this is clearly wrong — they can list many problems which the new paradigm does not solve. As defined by Kuhn [2], however, a paradigm does not solve existing problems, rather it provides a way of understanding and solving problems. A potential new paradigm will only be successful if it looks like it potentially provides better ways of solving more problems. A new paradigm may even introduce new problems which did not exist under the old paradigm, for example, for someone used to experience-based forensic comparison the data-based component of the new paradigm presents serious new practical problems. An example of a problem which opponents of the new paradigm use to argue against its adoption is as follows: There are differences between speakers in terms of language and dialect spoken and individual speakers differ in speech style from occasion to occasion, e.g., on one occasion they may speak calmly and on another occasion they may speak angrily. The differences between calm and angry speech in one dialect may be different from the differences between calm and angry speech in another dialect, and it is impractical to collect data on calm and angry speech in all languages and dialects. However, this describes a problem which exists and needs to be solved irrespective of the paradigm which has been adopted: In any data-based paradigm one must have data for the relevant language, dialect, and speaking styles, and in any experience-based paradigm one must have experience with the relevant language, dialect, and speaking styles; the cost involved in gathering these data or gaining this experience still has to be paid. Whether differences due to speaking styles are manifested differently in different dialects is irrelevant. If one is working on a particular case, then the case defines which combination of languages, dialects, and speaking styles are relevant, and if one is conducting general research in anticipation of potential future casework, then one will presumably decide which languages, dialects, and speaking styles to work on according to what one thinks will potentially be most useful in the future. The new paradigm actually makes it clear how one should proceed: One must collect audio recordings of speakers of the relevant language and dialect; for each speaker one must obtain at least one recording of them speaking in a calm voice and at least one recording of them speaking in an angry voice; one must build a forensic–voice-comparison system; and one must assess the reliability of this system on test data consisting of calm versus angry voice pairs. There are probably two possible ways of solving the problem, either by looking for acoustic properties which are robust to the speaking-style difference, or by building statistical models which can predict and compensate for differences in voice properties due to the speaking-style difference. The details of the possible solutions are not part of the paradigm, but the paradigm does provide a means by which to assess and decide which of the possible solutions is the best.

3. Conclusion

Based on my interpretation of the paradigm shift in forensic comparison science first described by Saks and Koehler [1], the new paradigm can be characterised as quantitative data-based implementation of the likelihood-ratio framework with quantitative evaluation of the reliability of the strengths of evidence produced. The new paradigm was widely adopted for forensic DNA comparison in the 1990s, and over approximately the last decade has begun to make inroads into the field of forensic voice comparison. There are outstanding problems for the implementation of the new paradigm, in particular the practical problem of collection and analysis of large databases of voice recordings. It will take a substantial investment of resources to solve these problems to the degree that forensic voice comparison in the new paradigm can become a practical everyday reality in many parts of the world. A great deal of money has been spent on the development of infrastructure and research for forensic

DNA profile comparison. I hope that funding agencies heed the call made by the US National Research Council [4] and provide adequate funding to develop other branches of forensic science including forensic voice comparison.

Acknowledgments

The writing of this paper was supported financially by Australian Research Council Discovery Grant No. DP0774115. Thanks to Didier Meuwley, Philp Rose, Yuko Kinoshita, Michael Jessen, Cuiling Zhang and two anonymous reviewers for discussion of ideas and comments on earlier drafts 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] T.S. Kuhn, The structure of scientific revolutions, University of Chicago Press Chicago, IL, 1962.
[3] T.S. Kuhn, The Structure of Scientific Revolutions. 2nd ed, University of Chicago Press Chicago, IL, 1970.
[4] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, National Academies Press, Washington, DC, 2009.
[5] C.G.G. Aitken, F. Taroni, Statistics and the evaluation of forensic evidence for forensic scientist, 2nd edWiley, Chichester, UK, 2004.
[6] I.W. Evett, The theory of interpreting scientific transfer evidence, Forensic Science Progress 4 (1990) 141–179.
[7] 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.
[8] D.J. Balding, Weight-of-evidence for forensic DNA profiles, Wiley, Chichester, UK, 2005.
[9] 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.
[10] I.W. Evett, Towards a uniform framework for reporting opinions in forensic science case-work, Science & Justice 38 (1998) 198–202.
[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] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication 31 (2000) 193–203.
[14] 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, doi:10.1016/j.csl.2005.08.005.
[15] 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, doi:10.1109/TASL.2007.902747.
[16] P. Rose, Forensic speaker identification, Taylor and Francis London, UK, 2002.
[17] P. Rose, Technical forensic speaker recognition, Computer Speech and Language 20 (2006) 159–191, doi:10.1016/j.csl.2005.07.003.
[18] L.A. Foreman, C. Champod, I.W. Evett, J.A. Lambert, S. Pope, Interpreting DNA evidence: a review, International Statistics Journal 71 (2003) 473.
[19] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A hierarchy of propositions: deciding which level to address in casework, Science & Justice 38 (1998) 231–239.
[20] P. Donnelly, Appealing statistics, Significance 2 (2005) 46–48.
[21] G.S. Morrison, Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework, Australian Journal of Forensic Sciences 41 (2) (2009) 1-7, doi:10.1080/00450610903147701.
[22] D. Meuwly, Forensic individualisation from biometric data, Science & Justice 38 (2006) 198–202.
[23] 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, doi:10.1558/ijsll.v14i1.137.
[24] N. Brümmer, L. Burget, J.H. Cernocký, O. Glembek, F. Grézl, M. Karafiát, D.A. van Leeuwen, P. Matejka, P. Schwarz, A. Strasheim, Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006, IEEE Transactions on Audio, Speech and Language Processing 15 (2007) 2072–2084, doi:10.1109/TASL.2007.902870.
[25] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Computer Speech and Language 20 (2006) 230–275, doi:10.1016/j.csl.2005.08.001.
[26] D.A. van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: C. Müller (Ed.), Speaker Classification I: Selected Projects, Springer-Verlag, Heidelberg, Germany, 2007, pp. 330–353.
[27] 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, doi:10.1121/1.3081384.
[28] G.S. Morrison, Y. Kinoshita, Automatic-type calibration of traditionally derived likelihood ratios: forensic analysis of Australian English /o/ formant trajectories, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association, 2008, pp. 1501–1504.
[29] D. Ramos Castro, Forensic evaluation of the evidence using automatic speaker recognition systems. PhD dissertation, Universidad Autónoma de Madrid, Madrid, Spain, 2007.
[30] M. Jessen, Forensic phonetics, Language and Linguistics Compass 2 (2008) 671–711, doi:10.1111/j.1749-818x.2008.00066.x.
[31] 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.
[32] 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.
[33] L.G. Kersta, Voiceprint identification, Nature 196 (1962) 1253–1257.
[34] J.S. Gruber, F. Poza, Voicegram Identification Evidence, vol. 54, American Jurisprudence Trials, Westlaw, 1995.
[35] L.M. Solan, P.M. Tiersma, Hearing voices: speaker identification in court, Hastings Law Journal 54 (2003) 373–435.
[36] R. Schwartz, Voiceprints in the United States — Why they won't go away, Proceedings of the International Association for Forensic Phonetics and Acoustics Conference, July 23–26, Göteborg, Sweden, 2006. [Retrieved September 2009 from: http://www.ling.gu.se/konferenser/iafpa2006/].
[37] T. Cambier-Langevald, Current methods in forensic speaker identification: results of a collaborative exercise, International Journal of Speech, Language and the Law 14 (2007) 223–243, doi:10.1558/ijsll.2007.14.2.223.
[38] J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC, Boca Raton, FL, 2005.
[39] 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.
[40] P. Rose, G.S. Morrison, A response to the UK position statement on forensic speaker comparison, International Journal of Speech, Language and the Law 16 (2009) 139–163, doi:10.1558/ijsll.v16i1.139.
[41] D.H. Kaye, G.F. Sensabaugh Jr., in: D.L. Faigman, M.J. Saks, J. Sanders, E.K. Cheng (Eds.), DNA Typing, Modern Scientific Evidence: The Law and Science of Expert Testimony, vol. 4, Thomson West, Eagan, MN, 2008, pp. 83–224.
[42] S.R. Lewis, Police applications of speech and tape recording analysis. Philosophy of speaker identification, Proceeding of the Institute of Acoustics 6 (1984) 69–77.
[43] 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.
[44] P. Rose, Identifying criminals by their voice: the emerging applied discipline of forensic phonetics, Australian Language Matters 5 (2) (1997) 6–7.
[45] 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.
[46] P. Rose, F. Clermont, A comparison of two acoustic methods for forensic speaker discrimination, Acoustics Australia 29 (2001) 31–35.
[47] 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.
[48] Q.Y. Kwan, Inference of Identity of Source, PhD dissertation, University of California, Berkeley, USA, 1977.
[49] 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.
[50] D. Meuwly, A. Drygajlo, Reconnaissance automatique de locuteurs en sciences forensiques: Modélisation de la variabilité intralocuteur et interlocuteur, Proceedings of 5eme Congres Français d'Acoustique, 2000, pp. 522–525.
[51] 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, Retrieved April 2009 from: http://www.ling.cam.ac.uk/francis/LawLang.doc.
[52] 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, D2-53–D2-54.
[53] 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.
[54] 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.
[55] 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.
[56] 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.
[57] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. PhD dissertation, University of Lausanne, Lausanne, Switzerland, 2001.
[58] 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, International Speech Communication Association, Toledo, Spain, 2004.
[59] 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, doi:10.1016/j.forsciint.2004.09.032.
[60] 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.
[61] A. Drygajlo, Forensic automatic speaker recognition. IEEE Signal Processing Magazine, (2007, March) 132–135.
[62] Y. Kinoshita, Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants, PhD dissertation, Australian National University, Canberra, Australia (2001).
[63] 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.
[64] 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.
[65] 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.
[66] G.S. Morrison, Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/, International Journal of Speech, Language and the Law 15 (2008) 247–264, doi:10.1558/ijsll.v15i2.249.
[67] 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.
[68] 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.
[69] T. Thiruvaran, E. Ambikairajah, J. Epps, FM features for automatic forensic speaker recognition, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association, 2008, pp. 1497–1500.
[70] T. Becker, M. Jessen, C. Grigoras, Forensic speaker verification using formant features and Gaussian mixture models, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association, 2008, pp. 1505–1508.
[71] D. Hodgson, A lawyer looks at Bayes' Theorem, The Australian Law Journal 76 (2002) 109–118.
[72] D. Hodgson, Speaker identification — a judicial perspective, paper presented at the Australian Research Council Human Communications Network Workshop on Forensic Speaker Recognition (FSI not CSI: Perspectives in State-of-the-Art Forensic Speaker Recognition), Sydney, New South Wales, Australia, 6–7 December 2007. [Retrieved September 2009 from: http://forensic-voice-comparison.net].
[73] D.V. Lindley, Probability and the law, The Statistician 26 (1977) 203–220.
[74] M. Coulthard, A. Johnson, An introduction to forensic linguistics: language in evidence, Routledge, London, UK, 2007.
[75] D. Watt, The identification of the individual through speech, in C. Llamas, D. Watt (Eds.), Language and Identities, Edinburgh University Press, Edinburgh, 2009, Ch. B2. [Pre-publication version retrieved 20 September 2009 from http://www-users.york.ac.uk/~dw539/watt2009.pdf].
[76] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defence attorney's fallacy, Law and Human Behaviour 11 (1987) 167–187.
[77] 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, Retrieved April 2009 from: http://www.lawcom.gov.uk/expert_evidence.htm.
[78] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Science and Justice 49 (2009) 161–164, doi:10.1016/j.scijus.2009.07.004.