# A response to the UK Position Statement on forensic speaker comparison

*Phil Rose & Geoffrey Stewart Morrison*

*School of Language Studies, Australian National University*

philip.rose@anu.edu.au
geoff.morrison@anu.edu.au

## 1 Introduction

A recent issue of the *International Journal of Speech Language and the Law* contained a “position statement concerning the use of impressionistic likelihood terms in forensic speaker comparison cases” (French and Harrison 2007). This position statement was the result of a collaborative exercise among a number of researchers and forensic practitioners working in the United Kingdom. The foreword states that:

> “the statement was circulated to all practising forensic speech scientists and interested academics within the UK. With one exception, all those contacted became co-signatories. The statement now reflects the all but unanimous position within the UK.” (p. 138)

The statement was also lodged with the prosecutorial bodies of Scotland, Northern Ireland, and England and Wales. For simplicity we will therefore refer to it as the *UK Position Statement*, with the proviso that it may not reflect the views of all interested parties in the UK[1], or have the force of law in any jurisdiction within the UK.

The editors of the *International Journal of Speech Language and the Law* invited responses to the UK Position Statement for publication in subsequent issues. This is one such response, a preliminary version of which was presented at the 17th meeting of the *International Association for Forensic Phonetics and Acoustics* in July 2008.

We first summarise the UK Position Statement as we understand it, and then present our response. It helps our exposition to precede our response with an outline of what we consider to be the correct framework for the presentation of forensic-voice-comparison evidence.

[1] Since we know of two “interested academics” in the UK who were not consulted, it is not the case that the UK Position Statement represents the views of all-bar-one of the “practising forensic speech scientists and interested academics within the UK”.

## 2 Description of the UK Position Statement

It is made clear in the foreword that the UK Position Statement refers to comparison of voice recordings performed by experts, and thus relates to *technical* and not *naïve* forensic voice comparison (for these terms see Nolan 1983: 7, 1997: 744–745).

### 2.1 Motivations and goals

In its foreword, the UK Position Statement notes that it was motivated by a concern about "the framework in which conclusions are typically expressed in forensic speaker comparison cases" (p. 137). The awareness of there having been a problem with the existing framework is said to have initially arisen from the Appeal Court of England and Wales ruling in *R v. Doheny and Adams* ([1996] EWCA Crim 728) which involved the prosecutor's fallacy related to the evidence-in-chief of a DNA expert.

The foreword to the UK Position Statement claims that it presents:

> "… [a] new approach [which] brings about a fundamental change in the role of the analyst and the evidence. In the past forensic speech scientists were often thought of as identifying speakers. Within the new approach they do not make identifications. Rather, their role becomes that of providing an assessment of whether the voice in the questioned recording fits the description of the suspect." (p. 138)

Footnote 2 of the UK Position Statement adds that the activity is therefore to be considered not identification, but comparison. The foreword also claims that the aim in developing the UK Position Statement was "… to bring the field [of forensic voice comparison] into line with modern thinking in other areas of forensic science" (p. 137), and that "This new framework is, at a conceptual level, identical to that used nowadays in the presentation of DNA evidence" (p. 138).

At the end of the document, the authors and signatories of UK Position Statement acknowledge that they :

> "… accept in principle the desirability of considering the task of speaker comparison in a likelihood ratio (including Bayesian) conceptual framework. However, [they] 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, §6)

It might seem that the UK Position Statement's rational eschewing of the likelihood-ratio-based framework in favour of its alternative proposal has a *faute de mieux* ring. However, this would be an unfair characterisation. It can be seen that attempts have clearly been made to manufacture a compromise with likelihood-ratio-based approaches, and it is thus possible to see this compromise – although it is not explicitly nominated as one – as a further motivation for the proposal. What we will argue is that it is not the best compromise, and that it is not conceptually equivalent to the framework for the presentation of DNA evidence.

## 2.2 The UK Framework

A flow chart of the framework proposed in the UK Position Statement (henceforth *UK Framework*) is presented in Figure 1. In the UK Framework, speech samples are to be compared in terms of two serially ordered factors: *consistency* and *distinctiveness*.

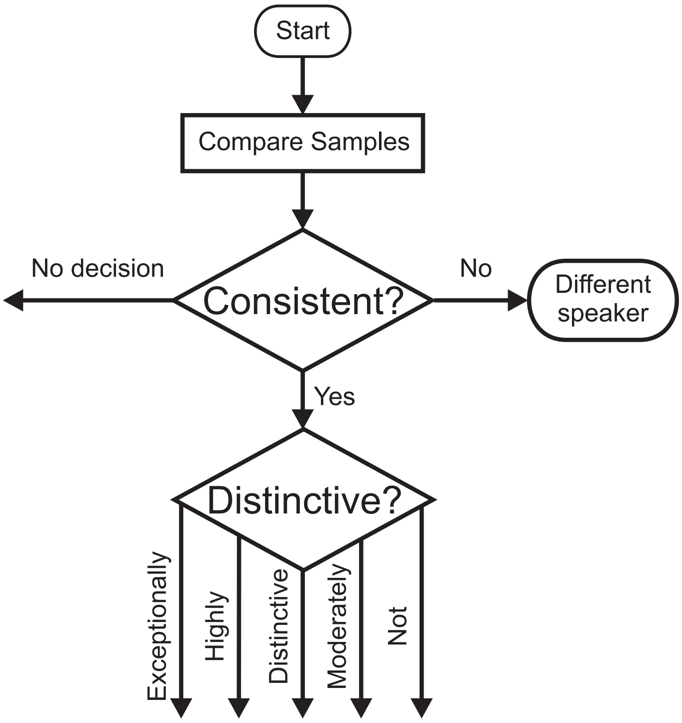

**Figure 1.** Flow chart representation of the UK Framework.

### *2.2.1 Consistency*

Consistency is characterised as “whether the known and questioned samples are compatible, or consistent, with having been produced by the same speaker” (p. 141, §4.1). It is assessed by “the degree to which observable features [are] similar or different” (p. 141, §4.1). Differences between known and questioned samples count against consistency unless “they can be explained by models of acoustic, phonetic or linguistic variation (e.g. by reference to differential channel characteristics, [or within-speaker] sociolinguistic, psychological and/or physical factors)” (p. 141, §4.1). Consistency is quantified on a three-point scale: *consistent*, *not-consistent*, or *no-decision*. If *not-consistent* is returned, the samples are declared to have been spoken by different speakers. If *consistent* is returned, one proceeds to consider the question of distinctiveness (thus consistency and distinctiveness are serially ordered, and a judgment on distinctiveness is only made if there has first been a positive determination on the issue of consistency). No instructions are provided as to actions following a *no-decision*.

### 2.2.2 Distinctiveness

The UK Position Statement emphasises that a positive determination of consistency does not imply that the known and questioned samples were necessarily spoken by the same person, since “the cluster of features leading to the consistency decision . . . [may] be shared by a substantial number of other people in the population” (p. 141, §4.2). It is implied that the likelihood that the samples have been produced by same speakers will be greater if their shared cluster of features is distinctive or unusual. Distinctiveness is assessed on a five-point scale ranging from *not-distinctive* to *exceptionally-distinctive*, the latter glossed with: “the possibility of this combination of features being shared by other speakers is considered to be remote” (p. 141, §4.2). There are no instructions as to how to proceed once a decision on distinctiveness has been reached. We presume that it is intended that the expert is to report that the samples are consistent with having been produced by the same speaker, and provide the determined degree of distinctiveness as an indicator of how unusual it would be to find this consistency if the two samples were not spoken by the same speaker.

## 3 Response to the UK Position Statement

First let us say that we applaud the motivation behind the UK Position Statement, and welcome its general direction. Specifically, we agree wholeheartedly with the goals of bringing the presentation of forensic-voice-comparison evidence into line with modern thinking in other areas of forensic science, and in particular of bringing it into line with modern practice in the evaluation of DNA evidence. These goals are not new: they were also proposed in Champod and Meuwly (2000), González-Rodríguez et al. (2006), González-Rodríguez et al. (2007), Rose (2002, 2003), Saks and Koehler (2005). We will argue, however, that the UK Position Statement fails to meet these goals.

Before our critical analysis of the UK Position Statement, it makes discussion easier if we first present what we consider to be the logically and legally correct framework for the evaluation of forensic comparison evidence: the likelihood-ratio framework. This is the framework which we believe represents the modern thinking in other areas of forensic science as exemplified by current practice in the evaluation of DNA evidence. Support for this position is provided by numerous textbooks, articles, and reviews written by forensic statisticians, legal experts, and forensic scientists, e.g. Aitken and Taroni (2004), Balding (2005), Champod and Meuwly (2000), Evett (1991, 1998), Friedman (1996), Good (1991), González-Rodríguez et al. (2006), Haigh (2005), Hodgson (2002), Lindley (1991), Robertson and Vignaux (1995).

### 3.1 The likelihood-ratio framework

The following account is abridged from Rose (2005: 49–54). It is written from the perspective that the objects of comparison are speech samples, but is in principle applicable to the evaluation of any type of forensic evidence (DNA, fingerprints, ballistics, toolmarks, etc.) where known and questioned samples are to be compared and it is possible to quantify physical properties which can vary across samples. For those interested in a fuller exposition of the likelihood-ratio framework, the following are recommended: Balding (2005), Lucy (2005), Robertson and Vignaux (1995), Rose (2002, 2003).

Typically in forensic voice comparison a recording of an unknown voice, usually of an offender, is to be compared with one or more recordings of a known voice, usually of a suspect or defendant. The interested parties (police/prosecution, trier-of-fact) want to know if the unknown (or questioned) voice comes from the same speaker as the known voice. They will usually understand that a definitive answer cannot be given; a trial is, after all, about making decisions in the face of uncertainty. So they will usually ask: *how probable is it that the samples have been said by the same person?* - a very reasonable way of putting it, since philosophers and statisticians will agree that the best way of quantifying uncertainty is by using probability (Lindley 1991). Implied, of course, will also be the rôle of evidence. That is, the question is really: *how probable is it, **given the voice evidence**, that the questioned and known samples have been said by the same person*? This is conventionally and conveniently formalised by the conditional probability expression at (1),

$$p(\mathrm{H}_{ss} \mid \mathrm{E}_{sp}) \tag{1}$$

where "$p$" stands for probability, "$\mathrm{H}_{ss}$" stands for the prosecution **H**ypothesis that the **same speaker** was involved, the vertical stroke "|" stands for "given", or "conditional upon", and "$\mathrm{E}_{sp}$" stands for the speech evidence - the inevitably present differences between the suspect and offender speech samples. It is usual to use odds instead of probability, so the formal representation becomes (2),

$$p(\mathrm{H}_{ss} \mid \mathrm{E}_{sp}) \,/\, p(\mathrm{H}_{ds} \mid \mathrm{E}_{sp}) \tag{2}$$

where "$\mathrm{H}_{ds}$" stands for the **H**ypothesis that the samples were spoken by **d**ifferent **s**peakers.

The solution to this equation is given by Bayes' Theorem, which has been known since at least the mid 1700s (Bayes, 1763). Bayes' Theorem is of paramount importance when one wants to know the probability of a hypothesis given the evidence, and this is what gives it its special status in forensic identification. Bayes' Theorem states informally that the probability of the hypothesis, given the evidence, can be estimated from two things: (1) how probable the hypothesis is, before the evidence is adduced; and (2) the strength of the evidence.

The odds form of Bayes' Theorem, applied to forensic voice comparison, is given at (3). It says that the odds in favour of it being the same speaker, given the speech evidence (this is what everyone wants to know and is called the posterior odds and is at the left of the equals sign), is the prior odds in favour of it being the same speaker times the strength of that evidence. Thus the odds in favour of it being the same speaker can be calculated from two terms: the *prior odds* and the *likelihood ratio*.

$$\underbrace{\frac{p\,(H_{ss} \mid E_{sp})}{p\,(H_{ds} \mid E_{sp})}}_{\textit{Posterior Odds}} = \underbrace{\frac{p\,(H_{ss})}{p\,(H_{ds})}}_{\textit{Prior Odds}} \times \underbrace{\frac{p\,(E_{sp} \mid H_{ss})}{p\,(E_{sp} \mid H_{ds})}}_{\textit{Likelihood Ratio}} \qquad (3)$$

The *prior odds* are the odds in favour of the hypothesis before the voice evidence is adduced. These are simply the probability that it is the same speaker divided by the probability that it is a different speaker. In its limit, it could be anyone in the world, but the prior odds can usually be considerably narrowed-down by taking into account obvious information in the voice like sex and accent, as well as other pragmatic information.

The *likelihood ratio* is the most important metric in forensic voice comparison because it is a measure of the *strength of the evidence* in favour of a hypothesis, and it is what the expert should try to estimate. The formula at (3) shows that the likelihood ratio too is a ratio of probabilities, but these probabilities are probabilities of *evidence*, not *hypotheses*. The likelihood ratio quantifies how much more likely you are to get the differences between the suspect and offender speech samples assuming they have come from the same speaker than assuming they have come from different speakers.

If you are more likely to get the speech evidence assuming that the samples came from the same speaker than from different speakers - if $p(E_{sp} \mid H_{ss})$ is greater than $p(E_{sp} \mid H_{ds})$ - that counts as support for the prosecution claim that the samples came from the same speaker. If, on the other hand, you are more likely to get the speech evidence assuming that the samples came from different speakers than from the same speaker - if $p(E_{sp} \mid H_{ds})$ is greater than $p(E_{sp} \mid H_{ss})$ - that counts as support for the defence claim. If you are just as likely to get the evidence assuming same-speaker as different-speaker provenance - if the ratio of $p(E_{sp} \mid H_{ss})$ to $p(E_{sp} \mid H_{ds})$ is one - the evidence is useless.

Thus the magnitude of the likelihood ratio quantifies the strength of the evidence: greater than unity means support for same-speaker claim; less than unity means support for different-speaker claim; unity (or values close to it) mean evidence is useless (or next to useless).

The main textbooks on the evaluation of forensic evidence (e.g. Robertson and Vignaux, 1995) and forensic statistics (e.g. Aitken and Stoney 1991, Aitken and Taroni 2004, Lucy 2005), stress that it is the role of the forensic expert to quantify the strength of the evidence by estimating its likelihood ratio: the probabilities of the evidence under competing prosecution and defence hypotheses:

> "The case made for this approach, whether the subject matter is DNA, glass fragments, clothing fibres or whatever, is overwhelming" (Haigh 2005)
>
> "Statistical evaluation, and particularly Bayesian methods such as the calculation of likelihood ratios . . . are the only demonstrably rational means of quantifying the value of evidence available at the moment: anything else is just intuition and guesswork." (Lucy 2005: 138)

### 3.2 Comparison, not identification

The UK Position Statement emphasises that its proposal is not to be considered identification, but comparison, and this is the term they use (and we have adopted). This choice is worth commenting on, because many terms are currently in use: *identification*, *recognition*, *verification*, and *discrimination*. In parts of the literature some of these terms, e.g. *identification* and *recognition*, have been used interchangeably, and in other parts different terms, e.g. *identification* and *verification*, have been used to designate logically different types of analysis or different applications (Rose 2002: Chapter 3). We agree with the UK Position Statement that *comparison* is the most appropriate term, but for slightly different reasons. Terms like *identification*, *verification,* and *recognition* imply the expression of a posterior probability (i.e., the probability that the suspect and offender voices are the same), and some have connotations of providing a categorical decision. As we argue below, and was stated in Rose (2002: 89), since it is logically not possible, and legally inappropriate, for a forensic expert to provide a posterior probability, there is no identification, verification or recognition involved, and we are therefore in agreement with the UK Position Statement that with respect to forensic work these terms should be eschewed in favour of a neutral term such as *comparison*, which does not carry connotations of providing a posterior probability decision. We would also suggest that since the objects of comparison are recordings of voices – voices are compared, speakers are not – we adopt the term *forensic voice comparison* rather than the UK Position Statement's *forensic speaker comparison*. Nolan (1983, 1996) gives the best current characterisation of a voice for forensic purposes. His semiotic account is also described in detail in Rose (2002: Ch. 10).

### 3.3 Prohibition on probability of hypothesis, given evidence

The first important proposal of the UK Position Statement is the recommendation that the expert refrain from giving the probability of hypothesis, given evidence [$p$(H|E)]. We strongly endorse this: for some time now this has indeed been the position adopted by forensic statisticians and more recently by some courts (see the historical discussion in Aitken and Taroni 2004: 108, 122–128, 153–155, 208–213, Balding 2005: 145–153). However, the UK Position Statement implies that the reason a forensic expert should not quote *p*(H|E) is because this gives a “false weighting” to the evidence which it relates to the prosecutor’s fallacy. The prosecutor’s fallacy refers to the erroneous transposing of evidence with hypothesis, i.e. replacing *p*(E|H) with *p*(H|E). This is the same as saying that because the evidence is 1000 times more likely under an assumption of guilt, the defendant is 1000 times more likely to be guilty (Aitken and Taroni 2004: 79–82, Balding 2005: 146–147, Donnelly 2005, Evett 1998, Thompson and Schumann 1987). Certainly forensic scientists should avoid making this error themselves, and should do everything within their power to prevent their testimony from being misinterpreted by lawyers, judges, and juries. But the UK Position Statement’s argument about “false weighting” does not address the real reasons why forensic scientists must provide the probability of evidence given hypothesis [*p*(E|H)], and why they cannot provide the probability of hypothesis given evidence [*p*(H|E)].

There are two reasons why the forensic expert cannot provide the probability of the hypothesis given the evidence: one logical, one legal. The logical reason follows trivially from Bayes’ Theorem. The posterior odds are determined by two things: the strength of the evidence (likelihood ratio) and the prior odds (see equation 3). The expert is not privy to the prior odds, hence a posterior cannot logically be quoted. The legal reason has to do with violations of the ultimate issue rule: In cases where the offender sample is truly incriminating, the expert’s pronouncement that the suspect is likely to have said the incriminating speech is equivalent to an expression of probable guilt, and this usurps the role of the trier-of-fact (the judge or the jury, depending on the legal system).

Although the UK Position Statement appears to condemn the practice of providing probability of hypothesis given evidence, there are two places where providing *p*(H|E) is in fact recommended. These are discussed within the following two subsections, one related to differences between DNA and speech data, the other with closed-set comparisons.

### *3.3.1 Differences between DNA and speech data*

The first violation of the prohibition on quoting *p*(H|E) occurs in relation to the *not-consistent* determination on the issue of consistency:

> “Where the samples are not consistent we see no logical flaw in making the statement that the samples are spoken by different speakers. This may be stated with a degree of confidence appropriate to the exigencies of the data.” (p. 141, §4.3)

Saying, with a given confidence, that the samples were spoken by different speakers because they are not consistent is a *p*(H|E) statement. *Pace* the UK Position Statement claim above, by Bayes’ Theorem it is, in fact, a logical flaw.

We suspect that this inconsistency has crept in because the authors of the UK Position Statement were attempting to adapt a model for DNA analysis without taking into account important differences in the nature of DNA versus speech evidence. Although the evaluation of forensic speech evidence can indeed be done in the same way as DNA, with likelihood ratios – this was demonstrated in a recent paper with both automatic and traditional approaches (González-Rodríguez et al. 2007) – one must be careful when drawing comparisons between DNA and forensic voice data. This is because of differences in the nature of the variation involved. The three aspects of variation which are of the greatest importance in forensics are the **type** of variation involved; how many **levels** of variation to account for; and the **magnitude** of the variation. DNA differs from speech in all three, but here it is the type and levels of variation that are of importance.

Variables can be either continuous or discrete, or a combination. DNA variables – typically the length of STR alleles at given loci - are discrete. With discrete variables it is possible to talk about a match, for example that both samples show a genotype having the same combination of 14, 16 at the D18 locus and 9.3, 9.3 at THO1 (Balding 2005: 3) [9.3, although it looks continuous, is not. It means that one of the 'repeats' only has three out of the expected four bases.] A non-match is also possible with DNA.

Whilst DNA evidence is discrete, speech evidence is continuous: cepstral coefficients, formant centre-frequencies, etc. are continuously valued variables, and even higher-level features such as the incidence of a particular allophone result in continuously valued proportions. Also, whereas the properties of speech vary from occasion to occasion, the DNA of a biological organism will be the same every time it is measured (making allowance for measurement error, contamination, somatic changes, transplants, chimera etc.).

Allowing for caveats, then, the properties of categoricality and invariance mean that if two DNA profiles do not match, the probability of getting this assuming that they have come from the same organism is zero. In this case the numerator of the likelihood ratio is zero and the posterior probability that they have come from the same organism, irrespective of the priors, is also zero. Allowing for caveats, DNA can therefore be used to provide definitive evidence of exclusion. Not so speech. In general, speech data do not, by their nature, allow such a definitive exclusion. We can imagine some conditions under which a voice comparison could result in a definitive exclusion, e.g., the vocal tract of a young child could not produce the lower formants of a typical adult male, but in such cases the voices are likely to sound so different that it would be highly unlikely that a forensic expert would be consulted.

Again allowing for caveats, given a match between two DNA profiles, the probability of observing this assuming that both samples come from the same organism is one (Aitken and Taroni 2004: 404, Evett 1998). With the numerator of the likelihood ratio unity, its magnitude is dependent on the size of its denominator. The denominator is the *random-match probability* (referred to in the

UK Position Statement as the *random occurrence ratio* [2]). This is the probability of getting a match with the obtained DNA profile if one randomly samples profiles from members of the relevant population. Since under such circumstances the likelihood ratio is equivalent to the inverse of the random-match probability, strength of DNA evidence can also be presented in the form of the random-match probability rather than the likelihood ratio.

It may be the case that an inappropriate transference from DNA analysis in the UK Position Statement has also resulted in problems involving the concept of random match. The UK Position Statement considers a case where, in the UK, a DNA match between the suspect and offender is established, with a random match probability of 1 in a million (i.e. one person in a million has a profile matching that of the offender), and that there are 60 million people in the UK. Under these circumstances the UK Position Statement quotes "… a one in sixty chance that the DNA came from the defendant" (p. 139). [3] The UK Position Statement continues:

> "The estimation that 1 person in a million will share the DNA profile is known as its 'random occurrence ratio'. Phoneticians can calculate the random occurrence ratio for very few features of speech. Exceptions are fundamental frequency (a measure of voice pitch), articulation rate (speed of speaking) and stammering." (p. 140, §3)

Since speech data are inherently continuous and it is a truism that a speaker never says exactly the same thing the same way twice, there is always variation between speech samples, and the numerator of a likelihood ratio derived from a forensic voice comparison can never be zero or one. The concept of random match is thus not applicable to continuously valued speech data. The strength-of-evidence from a forensic speaker comparison on continuously valued data can only be expressed in the form of a likelihood ratio.

Whereas random-match probability is certainly a meaningless concept with respect to inherently continuously-valued properties such as fundamental frequency,  arguably it may be possible to calculate random-match probabilities for incidence of speech features such as stammering. However, this would only be under the assumptions that a recording of someone who habitually stam-

[2] "The term 'random occurrence ratio' introduced by the court [in *R v. Dohney and Adams*] appears to be a synonym for match probability. This novel coinage is an unwelcome addition to the many terms already available: its unfamiliarity could confuse." Balding (2005: 152)

[3] The correct answer is actually a one in sixty-***one*** probability that the defendant is the source of the DNA trace (or ***odds*** of 60 to one against). Of the 60 million possible perpetrators in the UK, one is guilty and the remaining 59,999,999 are innocent. The guilty party will provide one match, and the 59,999,999 others will provide 60 matches (because the probability of a random match is 1/1,000,000, and 59,999,999 × (1/1,000,000) = 60 (to the nearest integer).  So there will be in total 61 possible matches. Out of this 61, one is the true match, and the rest are false positives, so the probability that the suspect left the trace is 1/61. A simplified formula for calculating the probability of guilt under these circumstances [$P(G|E) = 1/1+N*p$] is given by Balding (2005: 11), where $P(G|E)$ stands for the probability of ***G***uilt, given the ***E***vidence, $N$ is the number of people *other than the suspect* that could have been the perpetrator, and $p$ is the probability of a random match.

mers will always contain instances of stammering, and a recording of someone who does not generally stammer will never contain instances of stammers.

### 3.3.2 Closed-set comparisons

The second violation of the prohibition on giving *p*(H|E) occurs in section 5 of the UK Position Statement:

> "In a minority of cases, however, there is independent evidence (e.g. video surveillance) to show that a closed set of known speakers was present and participating in the conversation. In such cases the comparison task becomes an issue of who said what. In these circumstances, if the voices are sufficiently distinct from one another, we consider it justified to make categorical statements of identification." (p. 142, §5)

Making a categorical statement of identification, given sufficient distinctness, is a *probability of hypothesis, given evidence* statement, and a logical violation of Bayes' Theorem. Closed set comparisons can be treated in exactly the same way as open, from the point of view of the strength of evidence (see Rose 2002: 64, 74).

## 3.4 Two-stage assessment

Another important part of the proposal, and again a welcome step in the right direction, is its bipartite assessment in terms of consistency and distinctiveness. Section 4.2 of the UK Position Statement is correct in assuming that the value of the evidence is dependent not only on the *similarity* between the two samples, but also on how *typical* they are. Two very similar, yet typical, samples will not be valued as highly in terms of strength of evidence in favour of identity as two very similar, yet atypical, samples. This is a point not always understood, and it is not uncommon to encounter the assumption that identity follows from similarity alone. It is therefore good to see this made clear in the UK Position Statement.

At first glance the UK Framework's *consistency* and *distinctiveness* terms appear to parallel the *numerator* and *denominator* of the likelihood ratio discussed in section 3.1. However, the UK Framework's use of a bipartite assessment via *consistency* and *distinctiveness* is not equivalent to the calculation of a likelihood ratio. An essential feature of a likelihood ratio is that the numerator and denominator are measured on the same scale (they are both values from probability densities) and are directly associated with each other (i.e. in the form of a ratio). In the UK Framework, consistency and distinctiveness are serially ordered; are measured on different scales (one has three discrete levels and the other five); and they are not directly related to each other.

The trier-of-fact needs to know whether the differences between the speech samples are more likely to have arisen if they were spoken by the same speaker, or more likely to have arisen if they were spoken by different speakers, or equally likely to have arisen irrespective of whether they were spoken by the same speaker or by different speakers. It is not possible to do this unless both terms are quantified on the same scale and are directly related to each other. This we therefore consider a weakness of the bipartite UK Framework.

The UK Framework's two-stage analysis of *consistency* and *distinctiveness* is in fact reminiscent of Evett's (1977) evaluation of evidence in terms of *comparison* and *significance* stages. In this approach one first decides if there is a match on the basis of prior agreed criteria – say both samples lie within three standard deviations of each other. Then one assesses the probability of finding the observed degree of similarity in the relevant population (see criticism of this framework in Aitken and Taroni 2004: 10–11, and in Evett 1991). Although several variants of two-stage assessment have historically been applied to DNA evidence, this was superseded by likelihood-ratio evaluation, and two-stage approaches are not representative of modern practice (see Foreman et al. 2003, for a historical review of the interpretation of DNA evidence up to that point in time).

In addition to the systematic problems with the two-stage assessment of the UK Framework, there are problems with the structure of its component parts. Below we discuss the cliff-edge problem and multivariate data problem, both related to the fact that *consistency* and *distinctiveness* in the UK Framework have a finite number of categorical outcomes. We discuss these problems in relation to *consistency*, but it should be clear that near identical arguments can be made with respect to *distinctiveness*. The division into three versus five categorical outcomes is immaterial to the logic of the arguments. We also discuss the unfortunate choice of the word "consistent".

### *3.4.1 The cliff-edge effect*

It is a phonetic truism that there are always differences between speech samples, even if they come from the same speaker repeating the same thing only seconds apart on the same occasion. Furthermore, these differences will be gradient, not categorical, and as such the difference between the samples cannot be adequately captured by the UK Framework's ternary categorical outcome.

Consider vowel formants as typical gradient features. The situation often arises where a set of vowels from the suspect and offender are compared with respect to their formant centre-frequency distributions. Table 1 gives mean and standard-deviation values calculated from the formants of tokens of a single vowel phoneme taken from two intercepted telephone conversations. For the sake of argument let us assume that it has been determined that that the distributions of the features are sufficiently close to normal to warrant using the mean and standard deviation as appropriate representations of central tendency and dispersion.

**Table 1.** Means and standard deviations for formant centre frequencies measured in hertz for 15 tokens of Australian English /ɔ:/ in two different intercepted telephone conversations.

| | F2 | F3 |
|---|---|---|
| **Suspect** | | |
| mean | 1429 | 2298 |
| standard deviation | 30 | 67 |
| **Offender** | | |
| mean | 1450 | 2329 |
| standard deviation | 48 | 56 |
| **Difference between means** | 21 | 31 |

Consider F2 as a single feature. We imagine that, even in the absence of specific population data, on the basis of their experience most practitioners would agree that 21 Hz (a 1.5% difference) is well within the size of difference that might be expected for schwa F2 means from the same speaker talking normally on different occasions. If working in the UK Framework, they would return a decision of *consistent*. Since the UK Framework has not specified how one would arrive at a decision on consistency, we have to assume it would be by implicit reference to some quantification of variation such as the feature's standard deviation. Table 1 shows that for both the suspect and offender samples the standard deviation in F2 is larger than the difference in F2 means between the samples. But at what point does one change from deciding that the observed values of the feature are consistent to deciding that they are inconsistent with having come from the same speaker? Should the boundary be at two standard deviations, or perhaps three? Should one apply a frequentist statistical test, such as a *t*-test with a prescribed alpha-level of 0.05, or perhaps 0.01? Such an approach imposes a categoricality, with an attendant cliff-edge effect (Robertson and Vignaux 1995: 118). If the boundary were set at two standard deviations using the standard deviation from the suspect data, a difference in the F2 means of 59.9 Hz would be ruled *consistent* but a difference of 60.1 Hz would be ruled *not-consistent* (or would both be declared *no-decision*). Should the decision hang on a difference of 0.2 Hz? We contend that any metric of similarity for use with forensic speaker comparison should be a gradient not categorical.

### *3.4.2 Problems with multivariate data*

A further problem with the UK Framework's consistency factor arises from the necessity of comparing samples with respect to multiple features. The foreword to the UK Position Statement describes the process of forensic voice comparison as:

> "[involving] 'separating out' the samples into their constituent phonetic and acoustic 'strands' (e.g., voice quality, intonation, rhythm, tempo, articulation rate, consonant and vowel realisations) and analysing each one separately." (p. 138)

No problem there, since multidimensionality is one of the things that contributes to the forensic discriminability of voices. However, the Statement gives no indication of how ultimately to combine the individual evidence from each of these 'strands'. Returning to the formant data in Table 1 and the example of a two-standard-deviation boundary based on the standard deviation from the suspect data (the argument would apply equally well, irrespective of the features under consideration or of the prescribed threshold). What would be the decision with respect to consistency if the difference in the F2 means were 50 Hz, within the boundary, but the difference in the F3 means were 140 Hz, outside the boundary? What if a pair of samples were judged *consistent* on nine features but *not-consistent* on one? Again it is incumbent upon the UK Framework to specify how one would proceed under such circumstances. Here, as elsewhere, an example of the approach would have been useful, as it appears to us to be very difficult to implement. If one eschews categorical quantification, then one can simply use a multivariate gradient metric of similarity, such as the numerator of the multivariate likelihood ratio. The literature already contains many procedures for dealing with multivariate data within the likelihood-ratio framework (e.g. Aitken and Lucy 2004, Pigeon, Druyts, and Verlinde 2000, Reynolds, Quatieri, and Dunn 2000).

### *3.4.3 The semantics of 'consistent'*

Section 4.1 of the UK Framework defines consistency as "the degree to which observable features [are] similar or different" (p. 141, §4.1). This is certainly a coherent notion in that it is possible to estimate how similar the questioned voice is to the known voice in the specified feature(s). However, as it stands, consistency is epistemologically very weak and its implementation is problematic and far from clear. In particular, the choice of the word *consistency* to represent this parameter is not felicitous. In their book on the evaluation of evidence, Robertson and Vignaux strongly criticize its use by forensic experts:

> "Worst of all is the word 'consistent', a word in unfortunately common use by forensic scientists, pathologists and lawyers. … Unfortunately for clear communication … lawyers usually interpret 'consistent with' as meaning 'reasonably strongly supporting'." Robertson and Vignaux (1995: 56)

We suspect that juries, and perhaps even judges, are also likely to construe *consistent with coming from the same speaker* as meaning *likely to have come from the same speaker*, although we think that within the UK Framework it is clear that this is not what is intended by the term *consistent*. Robertson and Vignaux also point out, in its proper (though not usual) sense, *consistent* has little epistemological force, as "consistent with H" implies nothing about the likelihood that H is true (*the facts are consistent with H but unlikely to result from H* is coherent in this sense of *consistent*).

## 3.5 Problems with populations and samples

As stated above, the UK Position Statement says that a quantitative likelihood-ratio-based evaluation of evidence of the type outlined in section 3.1, however desirable, is not possible due to two factors: "problems of defining relevant reference populations", and "lack of demographic data" (p.142 §6). We readily acknowledge that these are real problems: probably the most pressing at the moment. The first is theoretical and relates to the choice of the relevant population to sample, and the size of that sample; the second is practical and relates to the actual collection of data. Below we address both in turn. We would argue that these problems, however real, do not prevent the use of likelihood ratios.

### *3.5.1 The appropriate reference population*

One of the problems for likelihood ratio-based approaches mentioned by the UK Position Statement concerns "defining relevant reference populations." In order to estimate the strength of evidence with a likelihood ratio a *reference*, or *background* sample from the relevant population is needed.[4] The UK Position Statement is certainly correct in assuming this is a problem, and not only for speech: it continues to be a challenge in the evaluation of DNA samples (Aitken 1991). The choice of the appropriate population to sample is strictly speaking dependent on the alternative hypothesis. In a typical case the prosecution will contend that the voice in the questioned sample is the same as that in the known sample, i.e., both are the voice of the defendant. The defence will contend that the voice in the questioned sample is not that of the defendant. However, the contention is unlikely to be simply that the speaker of the questioned sample is some other human being. Rather the claim is more likely to be implicitly that the voice is that of another speaker of the same sex as the defendant who speaks the same dialect of the same language. The defence hypothesis could be even more restricted: that the questioned voice is that of someone who (to a naïve listener such as a police officer or a lawyer) sounds like the defendant; or it could be that the source of the questioned voice is the defendant's brother; or even identical twin, should they have one (see the discussion in Lucy 2005: 129–133, with respect to population limits). It is likely that, at the time of conducting the analysis, the forensic-voice-comparison expert does not know the specifics of the defence hypothesis, and must therefore anticipate what it might be. Decisions on the relevant reference population will have to be made on a case-by-case basis (Evett and Weir 1998: 44–45; see also discussion in Aitken 1991).

A commonly asked question about reference population sampling, and one which may have been implied in the UK Position Statement's observations about populations, is: how big should the

[4] Unfortunately in the literature the terms *reference population* or *background population* have sometimes been used when what is actually meant is a *sample of the population*. A reference or background population [sic] is not the same as the population of all possible perpetrators (the latter may be adduced in the estimation of the prior odds).

sample be? This is a very important question, both from the point of view of research and in court (since it is one of the things that needs to be justified). The answer is that it depends on the precision required. Aitken (1991) discusses some approaches to solving this problem.

### *3.5.2 Lack of information on distribution*

By *lack of demographic data* we assume, on the basis of the following quote, that what is meant is a lack of knowledge of the distribution of typical forensic comparison features in the population:

> "for the overwhelming majority of voice and speech features examined in casework, it is simply not known how widely they are distributed in the population." (p. 140, §3)

The UK Position Statement echoes Rose (2002: 78) in pointing out the consequences of this - that if one doesn't know the incidence of a feature in the population it is not possible to quote the probability of observing it at random from that population, thus: "forensic phoneticians are unable to provide numerical statements of probability" (p. 140, §3).

Clearly it is true that if one does not have access to estimates of the distribution of speech properties in the relevant population then one cannot quantitatively estimate a likelihood ratio (with appropriate confidence limits/credible intervals) for a forensic voice comparison. However, by the same token it must also follow that without such estimates one also cannot make decisions with respect to the UK Framework's *distinctiveness* factor. It is not made clear how this inconsistency is to be resolved.

Perhaps the authors and signatories of the UK Position Statement have envisioned that *distinctiveness* judgments could be made by an expert based on their own experience, but then such an inference would also enable a likelihood ratio estimate: a feature value judged in the expert's opinion as "exceptionally distinctive" can also be characterised as having a low probability in the population (the denominator of the likelihood ratio).

Section 6 of the UK Position Statement concludes with:

> "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. In view of these difficulties, the framework we endorse is the one set out in 4 above." (p. 142, §6)

We must assume that "present context" should be read as "present context in the UK", although this is not made entirely clear. The immensely complex current linguistic situation in the UK is such that, given the identification of an accent in the offender and suspect samples, there is at present no, or next to no, information available on the distribution of (acoustic) features in that accent. Under these circumstances, neither quantitative likelihood ratio nor *distinctiveness* estimates are possible, unless the forensic expert goes and collects samples from the population. However, in Australia and

Spain forensic scientists have collected distribution data and presentation of forensic-voice-comparison evidence in the form of likelihood ratios has been tendered as reports and received by the courts. We acknowledge that this is facilitated by the fact that variation in Australian English accents appears to be much less than in English accents within the UK.

In the UK context the problem remains that if appropriate databases of speech recordings are not available, then the forensic expert will have to collect and analyse them. With a fully automatic system, the analysis is a relatively cheap operation, but for traditional acoustic-phonetic approaches a major investment in human labour is required to collect and analyse reference data. We note that further work is already underway in the UK on the collection of databases and measurement of features therein (Nolan et al. 2006, Hudson et al. 2007, Clermont et al. 2008); however, it looks as if the short-term reality in the UK will be that quantitative estimates of strength of evidence will not be possible in some, perhaps many, cases.

If a short-term solution is to present qualitative estimates of strength of evidence, as opposed to quantitative ones, then it would be better to make such statements in the form of a likelihood ratio, rather than using the UK Framework. Such an approach is recommended by Robertson and Vignaux, and Jessen, if population distribution data are not available:

> "To assess a likelihood ratio it is not essential to have precise numbers for each of the probabilities. The value of the evidence depends upon the ratio of these numbers. Therefore, if we believe that the evidence is 10 times more probable under one hypothesis that the other, the likelihood ratio is 10, whatever the precise values of the numerator and denominator may be. Often we will be able to assess this ratio roughly on the basis of our general knowledge and experience." Robertson and Vignaux (1995: 21)
>
> "Even in areas where no such population statistics exist, and therefore no quantification is possible, the Bayesian approach should be used as a conceptual framework that provides the logical backbone of voice comparison analysis." (Jessen 2008:13)

Clearly, many of the UK practitioners are highly competent phoneticians with extensive experience in comparing voices, forensically and otherwise, with respect to many features. They can be expected to have a good intuition about the expected magnitudes of both within-speaker and between-speaker differences in these features. Thus they could be expected to be able to proffer an expert opinion of the following kind: "From my experience I think you would be much more likely to get the differences I have listed between the offender and suspect speech samples assuming that they had come from the same speaker, rather than different speakers." (or *mutatis mutandis*). A qualitative statement like this, of the probability of evidence under competing hypotheses, surely would have some value to the court, and would also be consistent with our arguments for the likelihood ratio framework as the logically correct framework for the estimation of strength of evidence.

We are however in agreement with Lucy that a qualitative statement remains a poor substitute for a quantitative likelihood ratio (and its credible interval/confidence limits):

> "It would not be ideal, nor desirable, to use such an estimate to evaluate a key piece of evidence in a major criminal trial" (Lucy 2005: 137).

## 4 Summary & Conclusion

This response has critiqued the proposals outlined in the UK Position Statement for a change in approach to forensic speaker comparison. The authors and signatories of the UK Position Statement deserve credit for acknowledging the need for a change, and initiating one, in the UK. Specifically, we see as positive the Statement's recommendations to avoid the logically problematic traditional conclusions couched in terms of probability of hypothesis, given evidence; and to take into account both similarity and typicality of the speech samples under comparison. We also approve of the use of the term *comparison*, instead of *identification*, *verification*, or *recognition*.

We have identified three basic weaknesses in the proposal. Firstly, the apparent treatment of speech as if its features were discrete and invariant, like DNA, has the result that the approach would be very difficult to implement. Secondly, we pointed out the inconsistencies in first prohibiting but then allowing *probability of hypothesis, given evidence* statements; and proposing distinctiveness statements while acknowledging the absence of information upon which to base them. Thirdly, the absence of a way of relating the consistency and distinctiveness assessments does not help the trier-of-fact to interpret them.

Given the UK Position Statement's in-principle endorsement of the likelihood ratio approach, we have tried to criticise it on its own merits, and to avoid criticising it for not being a likelihood ratio framework. However, the bipartite assessment is not a likelihood ratio, and it is the use of likelihood ratios that characterises modern thinking on the evaluation of forensic comparison evidence. We would therefore reject the claim in the Statement's foreword that the UK Framework is "at a conceptual level, identical to that used nowadays in the presentation of DNA evidence" (p. 138). Consequently, we would also argue that the UK Position Statement has not achieved its goal, however laudable, of "… bring[ing] the field [of forensic voice comparison] into line with modern thinking in other areas of forensic science" (p. 137).

It will be interesting to see the proposed UK Framework implemented in research and casework. We, of course, would encourage forensic-voice-comparison researchers and practitioners world-wide to rapidly move towards adopting quantitative likelihood-ratio statements as standard (although there will be some features of forensic speech samples for which it will only ever be possible to give qualitative estimates). Given the amount of energy in the UK now going into acquiring quantitative data to underpin likelihood ratio-based forensic voice comparison, we hope this may actually soon be the case in the UK.

## References

Aitken, C. G. G. and Stoney, D. A. (1991) *The Use of Statistics in Forensic Science*. Chichester, UK: Ellis Horwood.

Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist* (2nd ed.). Chichester, UK: Wiley.

Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(4):109-122.

Balding, D. J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. Chichester, UK: Wiley.

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418.

Champod, C. and Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Communication*, 31: 193–203.

Clermont, F., French, J. P., Harrison, P., and Simpson, S. (2008) Population data for English Spoken in England. Paper presented at the 17th meeting of the International Association for Forensic Phonetics and Acoustics, Lausanne, Switzerland, July 2008.

Donnelly, P. (2005) Appealing statistics. *Significance*, 2(1): 46–48.

Evett, I. W. (1977) The interpretation of refractive index measures. *Forensic Science International*, 9: 209–217.

Evett, I. W. (1991) Interpretation: A personal odyssey. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 9–22. Chichester, UK: Ellis Horwood.

Evett, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3): 198–202.

Evett, I. W., and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sunderland, MA: Sinauer Associates.

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

French, J. P. & Harrison, P. (2007) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law*, 14(1): 137–144

Friedman, R.D. (1996) Assessing evidence. *Michigan Law Review*, 94: 1810–1838.

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., and Ortega-García, J. (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2-3): 331–355.

González-Rodríguez, J., Rose, P., Ramos, D., Torre, D., and Ortega-García, J. (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7): 2104–2115.

Good, I. J. (1991) Weight of evidence and the Bayesian likelihood ratio. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 85–106. Chichester, UK: Ellis Horwood.

Haigh, J. (2005) Review of statistics and the evaluation of evidence for forensic scientists. *Significance*, March: 40.

Hodgson, D. (2002) A Lawyer looks at Bayes' Theorem. *The Australian Law Journal*, 76: 109–118.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. (2007) F0 statistics for 100 young male speakers of standard Southern British English. In J. Trouvain and W. J. Barry (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences* 1809–1811.

Jessen, M. (2008) Forensic phonetics. *Language and Linguistics Compass*, 2(4): 671–711.

Lindley, D. V. (1991) Probability. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 27–50. Chichester, UK: Ellis Horwood.

Lucy, D. (2005) *Introduction to Statistics for Forensic Scientists*. Chichester, UK: John Wiley.

Nolan, F. (1983). *The phonetic Bases of Speaker Recognition*. Cambridge, UK: Cambridge University Press.

Nolan, F. (1996) Forensic phonetics. Notes distributed at the two-week course at the 1996 Australian Linguistics Institute, Australian National University, Canberra.

Nolan, F. (1997) Speaker recognition and forensic phonetics. In W. J. Hardcastle and J. Laver (eds) *The Handbook of Phonetic Sciences* 744–767. Oxford, UK: Blackwell.

Nolan F., McDougall, K de Jong, G., and Hudson, T. (2006) A Forensic Phonetic Study of 'Dynamic' Sources of Variability in Speech: The DyViS Project. In Warren & Watson (eds.) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* 13–18.

Pigeon, S., Druyts, P., and Verlinde, P. (2000) Applying logistic regression to the fusion of the NIST '99 1-speaker submissions. *Digital Signal Processing*, 10: 237–248.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10: 19–24.

Robertson, B. and Vignaux, G.A. (1995) *Interpreting Evidence*. Chichester, UK: Wiley.

Rose P (2002) *Forensic Speaker Identification*. London & New York: Taylor and Francis.

Rose P (2003) *The Technical Comparison of Forensic Voice Samples*. Issue 99, *Expert Evidence*. Freckelton I, Selby H, (series eds.). Sydney, Australia: Thomson Lawbook Company.

Rose (2005) Forensic Speaker Recognition at the beginning of the Twenty-First century. An overview and a demonstration. *Australian Journal of Forensic Sciences*, 37(2): 49–71.

Saks, M. J., and Koehler, J. J.  (2005) The coming paradigm shift in forensic identification science. *Science*, 309: 892–895.

Thompson, W. C., and Schumann, E. L. (1987) Interpretation of statistical evidence in criminal trials. The prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behaviour*, 11: 167-187.