Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Computer Speech and Language 20 (2006) 159–191

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Technical forensic speaker recognition: Evaluation, types and testing of evidence

Phil Rose [*]

Phonetics Laboratory, School of Language Studies, Australian National University,
Acton, Canberra, ACT 0200, Australia
Joseph Bell Centre for Forensic Statistics and Legal Reasoning, University of Edinburgh, Old College,
South Bridge, Edinburgh EH8 9YL, UK

Received 1 November 2004; received in revised form 29 July 2005; accepted 29 July 2005
Available online 1 September 2005

Abstract

Important aspects of Technical Forensic Speaker Recognition, particularly those associated with evidence, are exemplified and critically discussed, and comparisons drawn with generic Speaker Recognition. The centrality of the Likelihood Ratio of Bayes' theorem in correctly evaluating strength of forensic speech evidence is emphasised, as well as the many problems involved in its accurate estimation. It is pointed out that many different types of evidence are of use, both experimentally and forensically, in discriminating same-speaker from different-speaker speech samples, and some examples are given from real forensic case-work to illustrate the Likelihood Ratio-based approach. The extent to which Technical Forensic Speaker Recognition meets the *Daubert* requirement of testability is also discussed.

1. Introduction

Forensic Speaker Recognition (or Identification – the terms are used synonymously) is one of the most important, challenging, but perhaps least well understood applications of Speaker

[*] Tel.: +61 2 6125 4169.
E-mail address: philip.rose@anu.edu.au.

0885-2308/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2005.07.003

Recognition. There are several types (Rose, 2002, Chapter 5). When the decision is informed by theories and axioms from well established disciplines like Linguistics, Phonetics, Acoustics, Signal Processing and Statistics, the terms Technical Forensic Speaker Identification (Nolan, 1983, p. 7), or Forensic Speaker Identification by Expert (Broeders, 2001, p. 6) are often used. In contrast to this, so-called Naive Speaker Recognition refers to the unreflected everyday abilities of people to recognise voices. One important subtype of Naive Forensic Recognition (although its set-up and evaluation clearly requires the help of experts) occurs in voice line-ups (for a list of important references, see Rose, 2002, p. 106, for a description of a recent actual voice line-up, see Nolan, 2003).

Technical Forensic Speaker Recognition (TFSR) can be characterised with several, not necessarily orthogonal dichotomies, and the primacy of any particular dichotomy will naturally reflect the experience of the practitioner or laboratory in which TFSR is performed. Currently, probably the most important dichotomy – important because as will be shown below it has to do with the strength of evidence – is between the use of automatic speaker recognition methods and the use of more traditional approaches (although this paper will plead for a combination of both). Another possible distinction is in terms of logical task. Meuwly (2004a,b, pp. 11–12) describes a situation where TFSR can help an investigative executive – usually the police – by "establish[ing] a short list of the most relevant sources of a questioned recording among a set of known potential speakers". This use, clearly most akin to identification, tends to be associated more exclusively with automatic methods, which are thoroughly addressed by Gonzalez-Rodriguez et al. (this volume) and in the work of many other researchers in automatic speaker recognition. TFSR is, in the author's experience, far more commonly encountered in a sense akin to verification, where one or more samples of a known voice are compared with samples of unknown origin (Lewis, 1984, p. 69). The unknown samples are usually claimed to be of the individual alleged to have committed an offence, and the known voice belongs to the defendant or accused. The interested parties are then concerned with being able to say on the basis of the evidence whether the two samples have come from the same person, and thus be able either to identify the defendant as the offender, or exonerate them.

Another distinction can be drawn depending on whether the TFSR results are actually brought as evidence. In some laboratories, irrespective of the method used to compare voice samples, the requesting agency restricts the results to investigative purposes only and they are not the subject of expert testimony (Nakasone and Beck, 2001). Yet another distinction might be drawn in terms of whether there is a *known* sample or not, since sometimes an investigative executive wants to know whether two or more *unknown* samples come from the same speaker. And yet another distinction is whether TFSR refers to experimental activity – to test a particular research hypothesis perhaps – or whether it forms part of a real case.

Irrespective of the ways TFSR can be characterised, one thing remains central: evidence, and this paper will focus on three main topics related to evidence: the different types of evidence used in TFSR, the correct logical framework for the evaluation of that evidence, and the extent to which this evaluation can be tested to meet legal evidentiary standards. More detail may be found in Rose (2002, 2003).

2. Bayes' theorem and forensic identification

The post-1968 "new evidence scholarship" debate and the increased incidence, from 1985 onwards, of statistical evidence associated with forensic DNA profiling focussed attention on the proper evaluation of forensic evidence (Dawid, 2005, p. 6). As a result, practitioners in many different fields of forensic identification have become (or are becoming) aware of the fact that, however much the court or the police may desire otherwise, there are big problems associated with quoting the probability of the hypothesis given the forensic evidence (Aitken and Taroni, 2004; Robertson and Vignaux, 1995). Applied to TFSR this means that it will normally not be possible for an expert to say, for example, that they are 80% sure that the samples have come from the same speaker, given the similarities between them (Rose, 2002, 2003). Since it highlights the main difference between TFSR and most other applications of speaker recognition, where a binary decision is the usual desired outcome, it is important to rehearse the reasons why the forensic identification expert cannot quote the probability of the hypothesis given the evidence.

The court is faced with decision-making under uncertainty — in a case involving TFSR it wants to know how certain it is that the incriminating speech samples have come from the defendant. Probability can be shown to be the best measure of uncertainty (Lindley, 1991, pp. 28–30, 37–39). Therefore it is necessary to evaluate how much more likely the evidence – i.e., the differences/similarities between the speech samples – shows the defendant to have produced the incriminating samples than not to have produced them. This is shown by the ratio of conditional probabilities at (1), where H_{ss} = prosecution hypothesis that the samples were spoken by the same speaker; H_a = alternative (defence) hypothesis; E_{fsr} = forensic-speaker-recognition evidence adduced in support of H_{ss} (this evidence will be the similarities/differences between the offender and defendant speech samples); and $p(H_{ss}|E_{fsr})$, etc. stands for the probability that the same-speaker hypothesis is true, given the evidence

$$p(H_{ss}|E_{fsr})/p(H_a|E_{fsr}). \tag{1}$$

The solution to (1) is of course given by Bayes' theorem, and its centrality is the one non-negotiable thing in TFSR. The odds form of Bayes' theorem, again suitably subscripted to apply to the TFSR context, is given at (2). This formula has been styled "...*the* fundamental formula of forensic science interpretation" (Evett, 1998, p. 200).

$$\underbrace{\frac{p(H_{ss}|E_{fsr})}{p(H_a|E_{fsr})}}_{\text{Posterior odds}} = \underbrace{\frac{p(H_{ss})}{p(H_a)}}_{\text{Prior odds}} * \underbrace{\frac{p(E_{fsr}|H_{ss})}{p(E_{fsr}|H_a)}}_{\text{Likelihood Ratio}}. \tag{2}$$

As can be seen, (2) states that the posterior odds in favour of the hypothesis H_{ss} given the evidence E_{fsr} adduced in its support are the product of the prior odds in favour of the hypothesis and the likelihood ratio for the evidence. The Likelihood Ratio – the central notion in TFSR – is the ratio of the probability of getting the evidence assuming the hypothesis is true, to the probability of the evidence assuming an alternative hypothesis (one cannot estimate the probability of a hypothesis without comparing it to some alternative).

The prior odds are the odds in favour of the hypothesis before the evidence is adduced. Suppose the suspect is one of a group of five males known to be in a house at the time of an incriminating phone intercept. The prior odds are then 4 to 1 *against* them being the owner of the intercepted voice. Suppose further from comparison of known and unknown phone intercepts the evidence is estimated as 100 times more likely if the same speaker is involved (Likelihood Ratio = 100). The posterior odds on the suspect being the speaker now shift to (100 * 1/4 =) 25 to 1 in favour. The court must then interpret these odds – or more likely their corresponding probability. If it exceeds some previously determined value – beyond reasonable doubt or the balance of probabilities for example – the defendant is found by the court to have produced the speech samples. In this made-up case $O_{\text{post}}(\text{H}|\text{E}) = 25{:}1$, which corresponds to a probability of 25/26, or 96%. This is clearly beyond the balance of probabilities required in civil cases. Whether it constitutes *beyond reasonable doubt* is up to the court to decide (what a jury construes as beyond reasonable doubt often varies as a function of the perceived severity of the punishment).

Now, it is clear from Bayes' theorem that, unless the TFSR expert knows the prior odds, they logically cannot estimate the probability of the hypothesis. Since the TFSR expert is usually not privy to information that informs the prior odds – and in fact there are very good reasons why they should not be (Rose, 2002, p. 64, 74, 273–274) – they cannot logically state the probability of the hypothesis. Since this, in the author's experience, is precisely what is usually expected of the TFSR expert by just about everybody involved (instructing solicitors, councel, court and police), this can be a big problem (Boë 2000, p. 215; Rose 2002, pp. 76–78). It also needs to be acknowledged that this point is sometimes not appreciated even by the TFSR practitioners themselves, many of whom still formulate their conclusions in terms of $p(\text{H}|\text{E})$ (Broeders, 1999, p. 239). All of this may be related to the fact that, amply demonstrated in the early base rate neglect experiments like Tversky and Kahneman's "Cab" problem (Gigerenzer et al., 1989, pp. 214–219), people are disposed to ignore prior odds when asked to estimate the probability of a hypothesis given the evidence, and focus on the so-called diagnostic information (i.e., the Likelihood Ratio).

The main textbooks on the evaluation of forensic evidence, e.g., Robertson and Vignaux (1995), or forensic statistics, e.g., Aitken and Stoney (1991); Aitken and Taroni (2004), stress that it is the role of the identification expert to estimate the strength of the evidence by estimating its Likelihood Ratio – the probabilities of the evidence under competing prosecution and defence hypotheses. It is also possible to find this approach implemented in real case-work, both by experts and the judiciary. It is accepted in expert testimony involving DNA evidence for example, and here is an enlightened quote from a not so recent appeal court judgment in Doheny (1996, p. 8).

> When the scientist gives evidence it is important that he should not overstep the line which separates his province from that of the Jury... He will properly, on the basis of empirical statistical data, give the Jury the random occurrence ratio – the frequency with which the matching DNA characteristics are likely to be found in the population at large...
> The scientist should not be asked his opinion on the likelihood that it was the Defendant who left the crime stain, nor when giving evidence should he use terminology which may lead the Jury to believe that he is expressing such an opinion.

It would clearly be difficult to argue why TFSR practitioners should be exempt from this, and thus a correct format for a TFSR conclusion might go something like this. "There are always differences between speech samples, even from the same speaker. In this particular case, I estimate that you would be about 1000 times more likely to get the difference between the offender and suspect speech samples had they come from the same speaker than from different speakers. This, prior odds pending, gives moderately strong support to the prosecution hypothesis that the suspect said both samples." To which should probably be added, given our disposition to transpose the conditional (but at the risk of further confusion): "It is important to realise that this does not mean that the suspect is 1000 times more likely to have said both samples."

Quoting the Likelihood Ratio of the evidence, or using the Likelihood Ratio as a discriminant function, is often styled Bayesian, but it is of the utmost importance to realise that *the use of a Likelihood Ratio to help in evaluating the strength of evidence is not necessarily Bayesian in any special sense* (Hand and Yu, 2001, pp. 386–387). In formal statistics, the term 'Bayesian' implies, or is associated with, the use of subjective priors (Sprent, 1977, pp. 215–216). As just pointed out, legally the priors must not be the concern of the expert witness. Moreover, subjective priors can be anathema in the courtroom, if they ever get that far (Good, 2001, 5.5, 6.1, 6.2, 7). In Doheny (1996, p. 9) for example the ruling was "strongly endorsed" that "To introduce Bayes [sic] Theorem, or any similar method, into a criminal trial plunges the Jury into inappropriate and unnecessary realms of theory and complexity deflecting them from their proper task."

Although there are beginning to be signs of some positive cognisance of the appropriateness of Bayes' theorem on the part of the judiciary (e.g., Hodgson, 2002), it is nevertheless clear that a crucial distinction needs to be drawn between the forensic use of a Likelihood Ratio to quantify the strength of evidence and the additional use of subjective priors, and that the term 'Bayesian' is inappropriate when characterising the approach described in this paper. Since it is the use of a Likelihood Ratio which is crucial forensically, it would be obviously advisable to use a term something like 'Likelihood Ratio-based', rather than 'Bayesian', but I have followed current usage and persisted with 'Bayesian' in this paper.

It is not clear to what extent Bayesian approaches are being actually used in forensic speaker recognition. Gonzalez-Rodriguez et al. (2002, p. 173) say that the European Network of Forensic Science Institutes (ENFSI), for example, is engaging with Bayesian evaluation of evidence in the following fields: DNA, fibres, fingerprint, firearms, handwriting, tool marks, paint & glass, speech and audio. However, this is at least partially disputed by one of the reviewers of this paper from one of the biggest European laboratories who observed that ... "there are no ENFSI speech and audio labs that present their (non-automatic) identification results in Bayesian terminology." and that "...results are usually given in terms of subjective probabilities of the competing hypotheses", i.e., $o(\mathrm{H}|\mathrm{E})$.

The first published mention of the application of Bayes' theorem to TFSR occurred some 20 years ago, in Lewis (1984). The first real demonstration of the approach in automatic forensic speaker recognition research – stimulated by interaction between forensic and generic speaker recognition researchers[1] – occurred some fourteen years later (e.g., Meuwly et al., 1998). Since that pioneering work, as can be appreciated from Gonzalez-Rodriguez et al. (this volume); Meuwly (2001); Meuwly and Drygajlo (2001); Drygajlo et al. (2003), its use in automatic FSR has been well-established, where it is promoting worthwhile research which is making true progress. The use of Bayes' theorem in conjunction with more traditional approaches to TFSR was first mentioned in Rose (1997), and has been subsequently explored (in e.g., Rose, 1999; Kinoshita, 2001; Elliott, 2001; Rose et al., 2003; Alderman, 2004a,b).

Despite this relatively rapid evolution, Bayes is evidently taking some time to propagate, geographically and conceptually, in other FSR areas. For example, McDougall (2004, p. 116) states "In speaker identification, the phonetician needs to know the probability that speech samples from an unknown and a known speaker were produced by the same speaker,...". Currently

[1] I thank one of my reviewers for making this important point.

the most recent book on FSR, which contains no explanation whatsoever of how forensic speech evidence can be evaluated, nevertheless disarmingly proclaims: "Speech sound spectrography, sometimes called voice printing, provides investigators with accurate and reliable information about speaker identity." (Tanner and Tanner, 2004, p. 44). This is worrying, especially in a book that will be read and cited by Law professionals. It highlights well the continual need for cautionary reminders of the limitations of FSR like Boë (2000); Bonastre et al. (2003) and Ladefoged (2004).

3. Technical forensic speaker recognition and speaker recognition

The discussion above should have flagged that Technical Forensic Speaker Recognition and conventional, or generic Speaker Recognition (of the kind, say, that is evidenced in the NIST evaluations) are rather different. Meuwly (2004a,b), which are the source of the quotes in this section, brings their differences nicely into focus by situating them within the wider context of biometric technology, for which he first distinguishes two superordinate scenarios: "forensic" and "non-forensic", and then characterises each scenario with respect to several of their interrelated characteristics: in particular their aims and the methods used to achieve them. Much the same approach was used in Gonzalez-Rodriguez et al. (2002).

Meuwly's "non-forensic scenario" involves verification and identification. Its aim is to "Provide a binary decision on the identity of a human being" and "Minimise the errors". This contrasts sharply with the forensic scenario, which involves the various evidentiary, investigative and prosecution applications alluded to above, with an aim of "Quantify[ing] the contribution of the biometric trace material in the process of individualisation of a human being". The discussion above has shown how this is to work with speech – the "biometric trace material" is the speech available for comparison, and its contribution – to what extent it supports the hypothesis of same-speaker provenance – is quantified by a Likelihood Ratio. (In other words, in technical forensic speaker recognition, no recognition, verification or identification actually takes place, and to that extent the reference to recognition (or identification) in the name TFSR is a misnomer (Rose, 2002, pp. 87–90).) Both forensic and non-forensic scenarios involve binary decisions; null and alternative hypotheses; prior odds and thresholds, but differences in the nature and goal of the scenarios ensure that these components relate in different ways.

In generic speaker recognition, for example, the null hypothesis is that the test and reference samples have a common source, and the alternative hypothesis is that they are from a different source. In the forensic scenario, the null hypothesis – the prosecution hypothesis – is the same, but the alternative hypothesis – the defence hypothesis – does not have to be just that the samples have a different source.

In TFSR, quite often the alternative hypothesis H_a will simply be that the voice of the unknown speaker does not belong to the accused, but to another same-sex speaker of the language. This is often a default assumption, because under many jurisdictions there is no disclosure to a prosecution expert of H_a before trial. H_a may be that the offender voice is of someone who sounds like the accused (Rose, 2002, p. 65), or that the unknown speech is not from the accused but their brother. In the latter case, the logical evaluation is considerably simplified: the closed-set comparison means that the distribution of a set of features F in the suspect is compared with the distribution of F in one other person only (e.g., Rose, 2002, p. 256). An additional consideration is this. We might assume that there is probably a greater similarity between voices of siblings than between randomly chosen speakers, resulting in a bigger LR numerator, and a more difficult discrimination. However, there are some indications that, even though they may have similar vocal tract anatomy, siblings – especially identical twins – exploit the plasticity of the vocal tract and the nature of linguistic structure to *use* language differently. They may have different allophones for a phoneme, for example (Nolan and Oh, 1996; Rose, 2002, pp. 1–2), or habitually use different articulatory settings. Perhaps we see here the forensically much-neglected indexical function of language: speakers using language to signal identity.

The alternative hypothesis can on occasion get quite complicated. In a recent case, for example, it has been claimed, sensibly, that the questioned voice was not that of the female accused, but of a male speaker who sounds similar to the accused because her voice sounds like a male.

It is important to understand that the choice of the alternative hypothesis can substantially effect the estimate of the strength of the evidence. Fig. 1 shows, with DNA data (from Meuwly, 2005), the effect of different alternative hypotheses on the magnitude, and consequent probative value, of the estimated Likelihood Ratio. A situation is represented where the suspect's and offender's DNA have been compared using the Second Generation Multiplex Plus (SGM+) DNA profiling system, and a match declared. The SGM+ system compares alleles at ten different sites (D19, D3, D8, VWA, THO, D21, FGA, D16, D2, D18 – shown on the *y*-axis) together with a sex test. Results for the matches at the 10 loci are shown. The figures in brackets represent the genotype – the particular pair of alleles inherited from the parents observed at each locus (thus

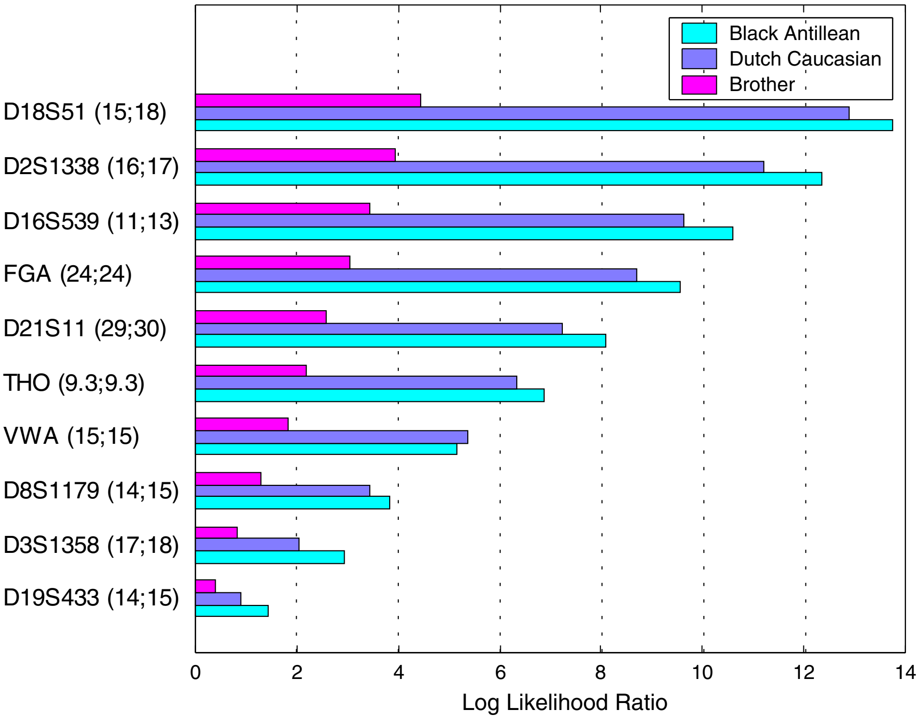

Fig. 1. Effect of different alternative hypotheses on the Likelihood Ratios from a DNA match (after Meuwly, 2005).

at locus D19 suspect and offender were both observed to have inherited 14 and 15 base repeats; at locus D3 they both had 17 and 18 repeats, etc.). The *x*-axis shows the cumulative magnitude of the estimated log Likelihood Ratio for the ten loci, under three different alternative hypotheses. The first alternative hypothesis is that the offender is a Black Antillean; the second is that the offender is a Dutch Caucasian; the third is that the offender is the suspect's brother.

The main thing to be seen in Fig. 1 is that the Likelihood Ratio estimate for the evidence – the match in DNA profile – changes depending on the alternative hypothesis. The difference is not much between the first two alternative hypotheses: if even only results from the first five loci can be taken into account the suspect is in trouble either way. But if the alternative hypothesis is that the suspect's brother was the donor, the value of the DNA match drops considerably, since there will be a much higher probability of shared genotype between siblings. For the first five loci, the match is only about 100 times more likely if the suspect were the donor rather than his brother. The limiting case, not shown in the figure, would of course be an alternative hypothesis that the donor was the suspect's identical twin (if he had one!). Then the value of the DNA evidence would be worthless, since the observed match would be equally possible under both prosecution and defence hypotheses.

The data in Fig. 1 can be used to make a further important point. Using Likelihood Ratios, evidence from different sources can be combined to give an overall Likelihood Ratio estimate for the totality of evidence in support of a hypothesis. In Fig. 1, the different sources are the matches at the different loci; in TFSR the different sources might be ten or so different phonetic or phonological features (Rose, 2002, pp. 60–61; 2003, pp. 3055–3059). Indeed Likelihood Ratios can be used to combine different *types* of evidence, for example TFSR evidence and blood-stain evidence. It can be appreciated from Fig. 1 that, although the magnitude of the estimated Likelihood Ratio may be small for a match at any one locus, it can get enormous when Likelihood Ratios from several loci are combined. This is because the loci are assumed to be independent (they are deliberately chosen to be on different chromosomes to maximise the probability of their independence) and therefore the overall Likelihood Ratio can be derived as the product of the Likelihood Ratios for the individual loci (Aitken and Stoney, 1991, p. 154; Robertson and Vignaux 1995, p. 166). Independence of features in TFSR, and hence their combination, is a problem – as is, to an extent, the assumption of independence of DNA features (Balding, 2005, pp. 20–21) – and is addressed later in this paper.

The assignment of priors is another way in which the two scenarios differ. In "non-forensic" discrimination the choice depends on the scenario – the cost of an error in classification, for example. Forensically, the prior is theoretically not subject to such determinism, and, as already pointed out, indeed may usually lie outside the expert's ken, and not be part of the their contribution at all. In some forensic areas, however, e.g., handwriting comparison, a prior of 0.5 is pragmatically assumed for both hypotheses, in order to allow an expert to quote a posterior probability to the court (Köller et al., 2004). When this happens it is made clear that the prior can be changed by the court at any time.[2]

Finally, it can also be appreciated that, strictly speaking, the nature of the Likelihood Ratio means that the threshold is fixed at 1 (or 0 for log-based quantification). In ASR, on the other

[2] I thank one of my reviewers for pointing this out and supplying the reference.

hand, the threshold is variable, and operationally determined by other factors like the equal error rate.

Thus it can be appreciated that, although the same components are often involved in forensic and non-forensic scenarios, they partition in different ways, depending on the scenario. A binary decision *is* involved in the forensic scenario, for example: between guilt and innocence (I ignore the possibility of the third verdict in Scotland). But this decision is the province of the court, not of the expert.

Perhaps the most important difference between the two scenarios relates to replicability. The notion of uniqueness is a salient characteristic of Forensic Speaker Recognition: ‘‘Forensic Scientists... must try to assess the value as evidence of single, possibly non-replicable items of information about specific hypotheses referring to an individual event’’ (Robertson and Vignaux, 1995, p. 201). Each case is unique. The evidence is unique, as well as, in principle, the alternative hypothesis. The prior will also be unique. This ubiquitous uniqueness guarantees non-replicability, a property which precludes the assessment of probability of guilt in frequentist terms (Lindley, 1991, p. 48, 49). This contrasts markedly with non-forensic scenarios, where replicability is an essential aspect, both experimentally and in real world application. In verification, for example, repeats of key utterances can be requested, and stored templates of subjects' voices can be retrieved as many times as necessary.

4. Likelihood ratio

The likelihood ratio (LR) is by far the most important construct in TFSI, since it quantifies the strength of the evidence in support of the hypothesis, according to the axiom of the Law of Likelihood (Royall, 2000, p. 760). Its numerator estimates the probability of getting the evidence assuming that the prosecution hypothesis is true; its denominator estimates the probability of the evidence under the alternative, defence, hypothesis. The relative strength of the evidence in support of the hypothesis is thus reflected in the magnitude of the LR. The more the LR deviates from one, the greater support for either prosecution (for $LR > 1$), or defence (for $LR < 1$). The more the LR approaches unity, the more probable is the evidence under both prosecution and defence hypotheses, and thus the more useless. Equivocal evidence tends to be a much underrated concept, since it is often assumed, in a binary forensic mindset, that for example if the prosecution hypothesis is not tenable, then the defence hypothesis must be true. The possibility of equivocal evidence as revealed by the LR shows that not only is one hypothesis useless – both are. So it is no good defence claiming that absence of evidence in support of the prosecution claim means automatic support for their position.

Verbal equivalents for LRs exist. Champod and Evett (2000, p. 240) proposed a set of terms for use at the British Forensic Science Service. For example, for $100 < LR < 1000$, evidence is described as giving ‘‘moderately strong’’ support for the prosecution hypothesis. However, neither the verbal equivalents nor their use is universal – for Royall (2000, p. 760), for example, LRs of 8 and 32 count as ‘‘fairly strong’’ and ‘‘very strong’’, respectively. Moreover, their use can be criticised as circular: in response to the claim that the evidence gives ‘‘strong support’’ to the hypothesis it can be enquired what is meant by ‘‘strong support’’, the only real response to which involves reference to the original LR (Rose, 2003, p. 2055).

There are other problems with the Likelihood Ratio and Bayesian evaluation of evidence. One is that it is difficult to come to terms with the idea that, for example, "strong support" is being claimed for a hypothesis which can be overturned when the prior odds are taken into account (although it is in fact sometimes the case that the prior odds are ignored by the court – whether by commission or omission is not clear). Also, and intriguingly from the point of view of linguistic semantics, the apparently glib English construction 'limited/strong evidence in support of x′ may not translate so trippingly into other languages. Broeders (2004), for example claims this is so for Dutch and German (partly because Dutch *bewijs*/German *Beweis* translate in English to both *evidence* and *proof*.)

Finally, *pace* Rose (2002, p. 76), and as conceded by Robertson, Buckelton and Dawid in their round-table discussion on the Bayesian evaluation of evidence (Robertson et al., 2005), at the moment, Bayesian inference is not easy for the court to understand, and Likelihood Ratios are all too easily transposed into probabilities of hypothesis given evidence. The prospects are sanguine, however, since it can be shown (Gigerenzer, 2002, pp. 40–44 *et pass*; Gigerenzer and Hoffrage, 1995; Pinker, 1997, pp. 343–351) that human minds are capable of Bayesian evaluation, provided that the wording is carefully chosen and refers to incidence ("out of 100 people, 3 will have this disease") rather than probability ("there is a 3% probability of this disease").

In TFSR, the LR numerator quantifies the degree of *similarity* between the offender and suspect samples, and its denominator quantifies the degree of *typicality* of the offender and suspect samples in the relevant population. Then the more similar the two samples are, the more likely they are to have come from the same speaker and the higher the ratio. But this must be balanced by their typicality: the more typical the samples, the more likely they are to have been taken at random from the population under consideration, and the lower the ratio. The value of the LR is thus an interplay between the two factors of similarity and typicality. Bayes' theorem makes it clear that both these factors are needed to evaluate identification evidence: it is a very common fallacy to ignore both base rate and typicality and assume that similarity is enough: that if two speech samples are similar that indicates common origin (how often do we hear the triumphal *gotcha* cry "it's a match!" in *Crime Scene Investigation*, or *Law and Order*?).

In non-automatic approaches, since voices are heavily multidimensional, it is possible, in theory, to calculate LRs for each separate feature examined and then combine them into an overall LR. The easy combination of LRs (at least it is easy if the evidence is independent) is one of the beauties of the Bayesian approach. The conditions upon $p(\mathrm{H})$ are actually more complicated (Bernado, 2001), and involve, for example, assumptions of how well the data are statistically modelled, and other background knowledge, in TFSR for example whether a suspect is known to be bilingual.

5. Likelihood ratio formulae

There are two different approaches to estimating a Likelihood Ratio; they can be characterised as (quasi-) empirical and (quasi-) analytic. The empirical approach is more common in automatic FSR, and involves number-crunching the distribution of the differences/distances involved. It is also possible to work with an analytically derived formula for a Likelihood Ratio. This kind of approach is encountered more often when comparison of forensic samples is in terms of traditional features, e.g., Alderman (2004a,b); Elliott (2002); Kinoshita (2001, 2002), Rose (2003, pp. 5107–5112).

As stated in the *locus classicus* for forensic LR derivation: "There can be no general recipe [for a LR formula], only the principle of calculating the [Bayes'] factor to assess the evidence is universal" (Lindley, 1977, p. 212). The reason why there cannot be a single LR formula is that the features in terms of which forensic comparison proceeds have different statistical properties, depending on what is being compared. The means of refractive indices of glass, for example, cannot be expected to distribute in the same way as means of formant centre-frequencies of vowels. A pane of glass; the friction-ridge patterns on a finger tip; sequences of junk DNA; bite marks; are not really very much like the acoustic and linguistic structure in the speech of one human speaker communicating with another.[3] Thus, in the proper forensic evaluation of differences between speech samples, LR formulae appropriate for speech have to be used, and different FSR features will require different formulae. It is a measure of the complexity of speech that truly appropriate LR formulae have not yet been derived, although, as will be demonstrated below, formulae which simplify one or more of the assumptions about the nature of speech – for example that features are normally distributed – appear to perform surprisingly well when discriminating same-speaker from different-speaker pairs. One such formula is given at (3), as an example

$$V \cong \frac{\tau}{a\sigma} \times \underbrace{\exp\left\{-\frac{(\bar{x}-\bar{y})^2}{2a^2\sigma^2}\right\}}_{\text{similarity term}} \times \underbrace{\exp\left\{-\frac{(w-\mu)^2}{2\tau^2} + \frac{(z-\mu)^2}{\tau^2}\right\}}_{\text{typicality term}} \tag{3}$$

$\bar{x}, \bar{y}$	means of offender and suspect samples;
μ	mean of reference sample;
σ	standard deviation of offender and suspect samples;
τ	standard deviation of reference sample;
z	$(\bar{x}+\bar{y})/2$;
w	$(m\bar{x}+n\bar{y})/(m+n)$;
m, n	number in offender, suspect samples;
a	$\sqrt{(1/m+1/n)}$.

The use of this formula, originally from Lindley (1977, p. 208), was demonstrated in the forensic comparison of refractory indices of glass fragments. It consists of three terms. The first, a variance ratio term, quantifies the ratio of between- to within-subject variance; the second, a similarity term, quantifies how similar the glass found on a suspect is to the window glass broken at the crime scene; the third, a typicality term, quantifies how typical the recovered and trace material are of the particular type of window broken (e.g., factory windows). The term V is equivalent to likelihood ratio; it might be thought of as standing for value of evidence.

To demonstrate the use of the formula in a forensic speaker comparison, assume that offender and suspect both have a Broad Australian accent, and that both offender and suspect samples contained four stressed utterances each of the word *hard* [had] in sentence-final position. Assume

[3] For discussions of that forensic chestnut, the differences between fingerprints and voiceprints, see Bolt et al. (1970); Rose (2003, pp. 4122–4123).

further that F2 was sampled in mid-vowel duration of all eight tokens of the word *hard*, yielding a mean and standard deviation F2 (Hz), respectively, of 1279, 30 for suspect, and 1284, 30 for offender. Given, according to Bernard (1967), a mean and standard deviation F2 (Hz) of 1367, 102 for /a/ in Male Broad Australian English *hard*, the formula at (3) estimates the LR at about 6. This means one would be about six times more likely to observe this difference assuming that the samples had come from the same rather than different speakers.

Ideally, four considerations have to be numerically incorporated in forensic LRs for speech: (1) the normality, or otherwise, of the distribution of the feature; (2) the equality, or otherwise, of the sample variances; (3) the levels of variance involved; and (4) the amount of correlation between features. To the extent these aspects are not, or inadequately, incorporated, the LR estimate will be inaccurate. These are briefly discussed below.

5.1. Normality

Some forensic speech features, for example cepstral coefficients, appear to be distributed normally, and can be adequately modelled by normal distributions. This is probably an unrealistic default assumption for speech, however, as indeed for many other modalities (Lindley, 1977, p. 211). For example, F2 in mid back rounded vowels like [ɔ] or [o] may not be normally distributed (Alderman, 2004a, p. 179). The formula at (3) assumes normality. For non-normality, various formulae with simple numerical integration can be used (Lindley, 1977, pp. 211–212), or a kernel density/GMM estimation. The formula at (4), from Aitken (1995, p. 188), estimates a LR using a gaussian kernel density model. Modelling non-normal distributions with kernel densities, or any other method of smoothing, is problematic and needs care. Automatic algorithms exist for the choice of smoothing coefficient (denoted λ in this paper), but it is often better to rely on the expert's subjective judgement from experience as to how they expect the variable to distribute (Aitken, 1995, pp. 185–186). One of the problems is that there are often rather different numbers of observations involved in the distributions to be modelled, which then require different choices of values for λ.

5.2. Equality of variances

The value of a LR is clearly dependent on the variances of variables in the two samples being compared. In speech, of course, variance is ubiquitous. It is expected that different speakers will have different variances for a given feature; and that the same speaker will differ in their variance on different occasions. There is thus both between- and within-speaker variation in variance, and this will therefore make any LR estimate assuming equal variances less accurate. Incorporating this into a LR formula is not straightforward: it can be seen that the otherwise rather complicated formula at (4) still assumes uniform within-subject variance.

5.3. Levels of variance

For forensic speech comparison at least three different levels of variance need to be modelled: between-speaker variance; within-speaker variance; between-session variance. Incorporating three levels of variance into a LR formula has only recently been attempted (e.g., Aitken et al., in press)

$$\mathrm{LR} = \frac{K\exp\{-\frac{(\bar{x}-\bar{y})^2}{2a^2\sigma^2}\}\sum_{i=1}^{k}\exp\{-\frac{(m+n)(w-z_i)^2}{2[\sigma^2+(m+n)s^2\lambda^2]}\}}{\sum_{i=1}^{k}\exp\{-\frac{m(\bar{x}-z_i)^2}{2(\sigma^2+ms^2\lambda^2)}\}\sum_{i=1}^{k}\exp\{-\frac{n(\bar{y}-z_i)^2}{2(\sigma^2+ns^2\lambda^2)}\}}, \tag{4}$$

where

$$K = \frac{k\sqrt{(m+n)}\sqrt{(\sigma^2+ms^2\lambda^2)}\sqrt{(\sigma^2+ns^2\lambda^2)}}{a\sigma\sqrt{(mn)}\sqrt{[\sigma^2+(m+n)s^2\lambda^2]}}$$

and

$\bar{x}, \bar{y}$ means of offender, suspect samples;
m, n number of observations in offender, suspect samples;
s^2 variance in reference population (between-speaker variance);
σ^2 within-speaker variance;
λ smoothing factor for kernel density estimate;
a $\sqrt{(1/m)+(1/n)}$;
w $(m\bar{x}+n\bar{y})/(m+n)$;
k number of kernel functions;
z_i value at which probability density is evaluated for the *i*th kernel.

5.4. Feature correlation

"…the assumption of independence [of predictor variables] is clearly almost always wrong (naturally occurring covariance matrices are rarely diagonal)…" (Hand and Yu, 2001, p. 387). In speech, many features are correlated. For example, one would expect F2 and F3 centre-frequencies in non-low front vowels (e.g., [i ɪ e]) to be correlated, and massive correlation has been found between cepstral coefficients (Rose et al., 2004). This correlation needs to be taken into account when estimating a LR. It would clearly be wrong to estimate a separate LR for F2 and F3 in [i], for example, and then derive an overall LR from their product.

Likelihood ratios have been derived for the comparison of trace evidence (elemental ratios in glass fragments), which take into account correlation between variables (Aitken and Lucy, 2004; Aitken et al., in press), but as yet little work has been done on speech material. Interestingly, an experiment to test the discriminant performance of the approach of Aitken and Lucy (2004) on speech found that it did not perform quite as well as a Naive-Bayes (also called "Idiot's-" or "Independence-Bayes") approach which assumed, quite against indications, that all variables were independent (Rose et al., 2004). It is apparently not unusual for approaches which use a naive Bayes classifier to outperform competitors in this way. Reasons for this are explored in Hand and Yu (2001) and Rish (2001). However, the fact that one can obtain better Likelihood Ratio-based discrimination results by ignoring correlation between predictor variables is a problem. This is firstly because it is then not clear which LR to present in evidence: the more accurate one that takes correlation into account, or the one which ignores correlations but has a greater discrimination potential? It is a problem also because discrimination performance is usually our only method of demonstrating the reliability of Likelihood Ratio estimation in real-world cases.

6. Background data

The similarity between the forensic samples has to be evaluated for typicality against background (also called reference) data. The background data depends on the alternative hypothesis H_a, which needs careful consideration. If H_a is that the incriminating speech came from some other speaker, a representative distribution of the parameter for appropriately sexed speakers of that language is needed. If H_a is that the speaker is someone else with a similar-sounding voice, then ideally a distribution of the parameter in pairs of similar-sounding voices needs to be used.

Proper implementation of the LR-based approach requires that an adequate background distribution exists. In most cases – at least for traditional features – it does not, and its estimation can only be very approximate. In the three examples of real-world LR comparison to be given below the background distribution will be seen to be defective in at least two respects. In two comparisons the distribution is likely to have been estimated on too few subjects; in one comparison the number of subjects is probably sufficient, but the variable modelled is not quite the same (the actual variable being compared is the mean F2 centre frequency in /ɐ/ before /k/; the background distribution is of /ɐ/ F2 before /alveolar stops/). The lack of adequate background data is one of main factors that makes the accurate estimation of Likelihood Ratios problematic. In such cases it is advisable, especially from the court's point of view, to run so-called sensitivity tests (Good, 2001, Chapter 9, Section 3.1), and use parameters varying over an expected range to estimate a range of LRs, rather than a single LR.

7. Evidence and forensic speaker recognition features

It is necessary to distinguish three different things when discussing the notion of strength of forensic evidence as quantified by a LR. Firstly, there is the raw data: for example a fingerprint, a bite mark, blood spatter, an analog recording of speech on a cassette or a digitised speech sample on a CD. Next there is information that the court receives from the expert witness concerning their qualifications, experience, methods of analysis, and findings: this is evidence in the legal sense: relevant information that the court has then to weigh. Finally, there is the evidence in the Bayesian sense – information that the expert witness extracts from the raw data, quantifies and uses in the LR estimate. In TFSR, this kind of evidence is then the ensemble of differences between the forensic speech samples when extracted and quantified with some analytic technique, such as formant centre-frequencies, cepstral coefficients or classical phonemic analysis.

It is important to note these distinctions, because, firstly, typically there will be information in the raw data that is not exploited. This will be due, trivially, to time constraints, but much more importantly also to analytic approach: a local, perhaps formant-based approach will be unable to make use of much of the individual-specific information in the samples that can be extracted automatically; a global automatic approach is by definition unlikely to pick up potentially crucial between-sample differences in the realisation of a single phoneme. It is also important to remember that, as with other areas of forensic science, different methods can result in different strengths of evidence, even on the same raw data.

7.1. Types of features

There are four main types of Bayesian evidence in FSR, usefully (but not crucially) characterised as the intersection of two binary features: *Auditory*/*Acoustic* and *Linguistic*/*Non-linguistic* Rose (2002, pp. 34–40).

7.1.1. Auditory features

Auditory features are those that can be extracted by trained, theoretically-informed listening. The theory is informed by all aspects of linguistic structure, not just phonetics, and the training is the kind provided by tertiary-level courses which teach (1) how to reliably transcribe and productionally interpret any speech-sound (and ideally any human vocalisation), and (2) how to analyse linguistic structure and the way it varies, both between- and within-speakers. An auditory analysis is precisely that – analytic – and not a holistic, undifferentiated and unreflected "these two samples sound to me as if they have come from the same speaker", (although it is in principle possible to assign a Likelihood Ratio to natural gut feelings like this (Rose, 2003, pp. 3061–3062)).

7.1.2. Acoustic features

Acoustic features are self-explanatory, and can be subcategorised into *traditional* and *automatic*. Traditional features relate in a direct way to aspects of speech production, like formant centre-frequencies, F0, or jitter. Automatic features are those like cepstral, or delta-cepstral coefficients. One is tempted to say that the choice between traditional and automatic features represents the most basic dichotomy within FSR, since many other methodological differences covary with them. The distinction between traditional and automatic features is important, since it reflects a tension between interpretability and discriminant power: traditional features have much greater interpretability – more *Anschaulichkeit* – which is a bonus for explanations and justifying methodology in court. Automatic features, on the other hand, are very much more powerful as evidence: they will, on average, yield likelihood ratios that deviate much more from unity (Rose, 2003, pp. 4095–4098). To demonstrate this important point, Fig. 2, from Rose et al. (2003) contrasts probability density distributions of log LRs calculated using traditional parameters (formant centre-frequencies) with LRs calculated with automatic parameters (cepstral coefficients). The data is the same in both cases: 240 same-speaker and ca. 28,000 different-speaker trials using non-contemporaneous Japanese telephone speech. It can be seen that the distribution for the LRs estimated from cepstral coefficients lies much further away from the threshold than the formant-based LRs, at least for the different-speaker comparisons (the probability of observing $LR < 1$ in different speaker trials was 99.96 with cepstral coefficients, but 92.0 with formants). It was found that analyses with both types of feature yielded useful strengths of evidence, but, given that the same-speaker resolution was fairly similar (see Fig. 2) the automatic approach, not surprisingly, was stronger on average by a factor of 18. With formants, a Likelihood Ratio bigger than unity

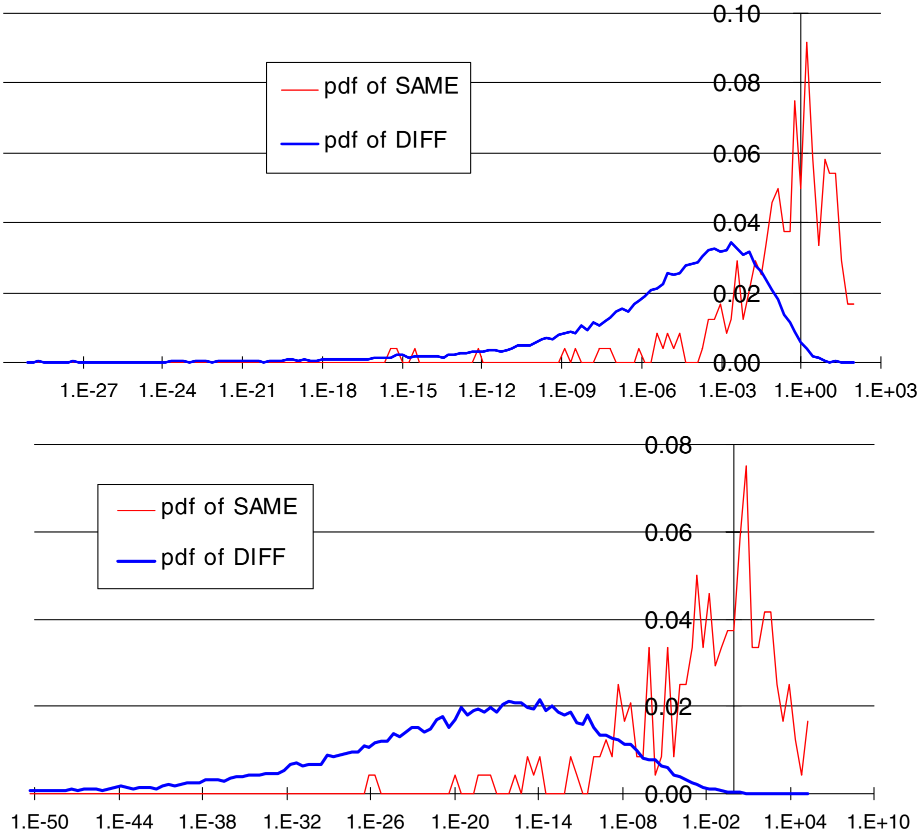

Fig. 2. Probability density distributions of log LRs for the comparison of 240 same-speaker (SAME) and ca. 28,000 different-speaker (DIFF) samples. Top = comparison using formants; bottom = comparison using cepstral coefficients. Horizontal axis shows LR value; vertical axis shows probability density. Vertical line shows location of LR = 0 threshold.

was on average about 50 times more likely if the samples were from the same speaker; with the cepstrum, LR > 1 was about 900 times more likely.

Although the particular disciplinary background of an expert will tend to influence their choice between automatic and traditional features, there is no reason why both types of features should not be combined in case-work (Rose, 2003, p. 193; Künzel et al., 2003) – especially since ease of combination of different types of evidence is one of the clear advantages of the Bayesian approach. Since different types of evidence are generally tapped by the two approaches, this would result in potentially even more powerful, and presumably more accurate, LRs.

7.1.3. Auditory vs acoustic features

Since there is evidence that the exclusive use of auditory or acoustic features is associated with considerable shortcomings, the consensus among practitioners is that both are necessary to evaluate differences between samples. An auditory approach on its own is problematic because it is possible, due to aspects of the resolution of the perceptual mechanism, for two speech samples to sound similar even though there are considerable acoustic differences between them (Nolan, 1990). By the same token, two forensic samples can have very similar acoustics and yet crucially differ in a single auditory feature. For example, one sample may uniformly have a labio-velar approximant [ʋ] for the English rhotic phoneme /r/, while the other is uniformly post-alveolar [ɹ] (Nolan and Oh, 1996; Rose, 2002, pp. 1–2).

There is often an enormous amount of potentially useful – even crucial – information available from the auditory features, although the evidentiary value of a feature is often language-dependent. For example, creaky phonation is a normal speech sound in Standard Vietnamese, and therefore of no forensic use; by contrast, it can be a marker of individuality in varieties of English, although even there its forensic use is restricted because it can function paralinguistically to signal temporary boredom, and linguistically to signal end of turn at talk.

Trivially, a prior auditory analysis is necessary to decide whether the samples are comparable in the first place, and if they are, what is to be compared – do we include emotional speech? laughter? screams? coughs? (cf French and Harrison, 2004; Yarmey, 2004). Auditory analysis is also needed for deciding how many speakers are involved, and partitioning the speech into putative speakers, since forensic speech samples are usually not monologues. It is also sometimes the case that during a conversation a questioned speaker is either identified by name by their interlocutor, or refers to themselves by name. It is then doubtful whether any further analysis – acoustic or auditory – is necessary to identify them, although such instances of meta-identification can provide very useful known reference data for estimating the within-speaker distribution of variables (which is a problem, whichever approach is used).

7.1.4. Linguistic and non-linguistic features

Linguistic features have to do with how the units of Language – the supremely human code that links speech sound to meaning – are organised and realised. Linguistic features can be broadly grouped into: phonological (having to do with speech sounds – e.g., the choice of /rum/ or /rʊm/ for *room*); morphological (having to do with the structure of words – e.g., the choice of /juθs/ or /juðz/ for the plural of *youth*); and syntactic (the ways words are strung together to form larger units like phrases or sentences – e.g., *I would have rathered to work* vs. *I would rather have worked vs. I rather would have worked*).

Speakers of the same language can and do differ in linguistic features, although this depends on the language. Samples in languages with a strong norm, and less dialectal variation, like Australian English, generally contain less such features. Samples in languages with less well established, or less prestigious norms, and extensive dialectal variation, like Chinese, generally contain more.

Non-linguistic features can be defined negatively as what is left when the linguistic ones are removed. These may be habitual articulatory or phonatory settings like the use of nasalised or breathy or creaky voice; lower than average pitch; fast or slow speech rates; etc. They may also be pathological features.

8. Examples of forensic application

8.1. Acoustic–linguistic features

One of the commonest acoustic–linguistic features used in forensic comparison is vocalic formant centre-frequencies. F1 (except possibly for low vowels) and F4 (except possibly for rhotics) are counter-indicated because of differential effects of the telephone transmission (Rose and Simmons, 1996; Künzel, 2001; Byrne and Foulkes, 2004), but F2 and F3 are usually reliably and usefully quantifiable for some vowels in even average quality recordings (Rose, 2003, pp. 5101–5113). As an example from case-work, Fig. 3 shows the mean F-pattern for 17 tokens of *yeah* [jə̟̞ː] said by the suspect during a police interview (suspects often say very little more than this) with the grand mean F-pattern of 15 of the suspect's *yeah*s from six telephone conversations intercepted about a year earlier. (The F-pattern was sampled as a function of equalised duration of the nucleus.) It can be seen that there is fairly good agreement between the mean time-normalised course of F2 and F3, but that the phone F1 is higher than in the interview, and the phone F4 is considerably lower. These are well-known effects of telephone transmission.

Features like formant centre-frequencies can be considered as linguistic because, due to the long-known relationship between the lower formants and auditory vowel quality (height, backness, rounding), the lower formants relate clearly to the linguistic unit being signalled. Also, of course, languages and dialects are known to differ in (normalised) lower vocalic formant frequencies.

Fig. 4 represents the evaluation of evidence in a fragment of case-work based on the F2 centre frequency of the second diphthongal target in /ɛɪ/ in the Australian English word *okay* (Rose, 2003, pp. 4119–4122). *Okay* is a very common word in conversations, and yields several forensically useful features. This particular frequency reflects how high and how front the speaker locates

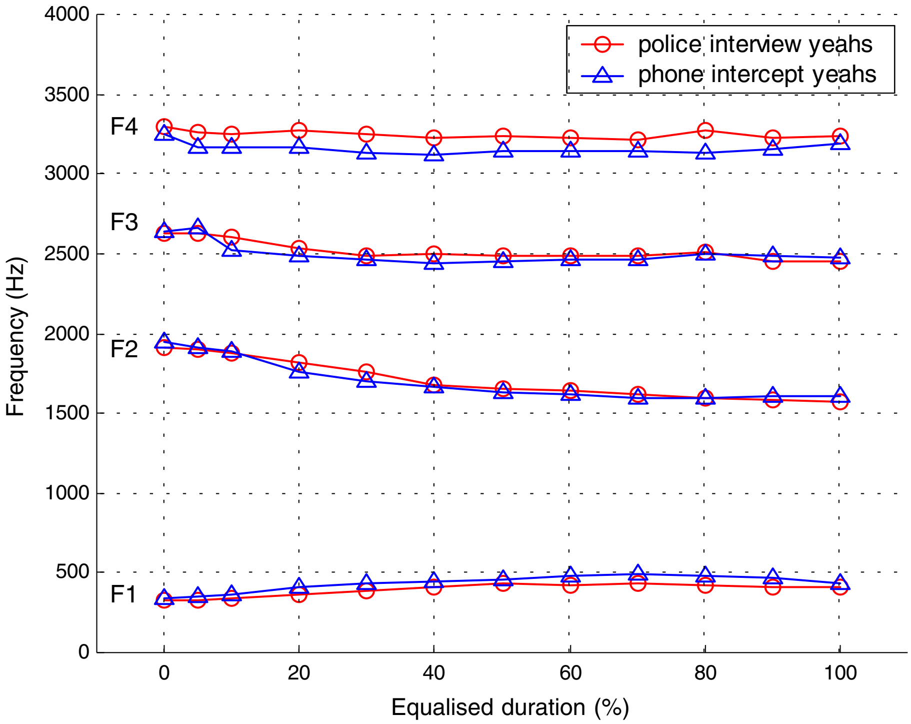

Fig. 3. Mean F-pattern for suspect's *yeah* during police interview compared with his grand mean F-pattern from known telephone intercept *yeah*s.

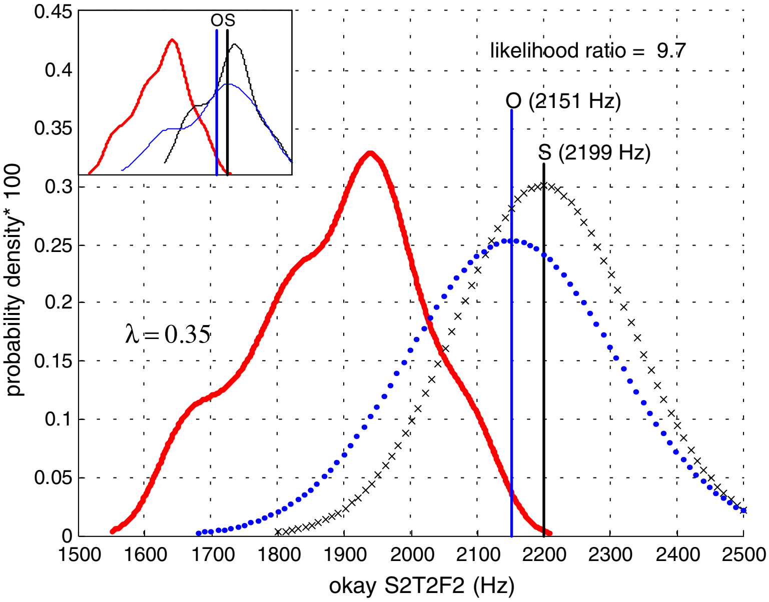

Fig. 4. Forensic kernel density estimation of an acoustic–linguistic feature in *okay*. Thick line = kernel density estimate of reference distribution. Offender and suspect sample distributions (dots, crosses) are modelled normally. O = location of mean of offender samples, S = location of grand mean of suspect samples. Insert shows kernel density distributions of offender ($\lambda = 0.75$) and suspect ($\lambda = 0.5$) samples.

their tongue body at the end of the diphthong, as well, of course, as the overall dimensions of their tract. In this particular case both suspect and offender samples were perceived to have a very close, very front offset to the /ɛɪ/ diphthong in this word. In Fig. 4, a comparison is shown between the mean value of 2151 Hz from four offender *okay*s in a single conversation, and a grand mean value of 2199 Hz from the means of several *okay*s in seven different known conversations of the suspect.

The difference between the suspect and offender means was evaluated using the kernel density estimation formula at (4) against the reference distribution of the same feature in the conversational speech of 10 male speakers of Australian English derived from Elliott (2002). In Fig. 4 the reference distribution is shown modelled with a Gaussian kernel density, and is mildly negatively skewed. The distributions of the offender and suspect observations are shown modelled normally in the main part of the figure, and modelled as Gaussian kernel densities, with different smoothing parameters, in the insert.

It can be seen in Fig. 4 that the probability density of the offender mean assuming it has come from the suspect, and the probability density of the suspect mean assuming it has come from the offender are fairly similar, compared to the probability density of both relative to the reference distribution. The ratio of similarity to typicality in this case appears therefore quite big. (The Fig. 4 insert shows that the degree of similarity will be slightly bigger if the distributions are modelled with kernel densities.) Nevertheless, the likelihood ratio is also of course a function of the variances involved, and it can be seen that, despite the fact that this feature tends to show a relatively large ratio of between- to within-speaker variance (Elliott, 2001) the standard deviation of the offender and suspect samples is about the same as the spread of the reference sample. This will have the effect of scaling the likelihood ratio down. The likelihood ratio in this case is 9.7: one would be about 10 times more likely to observe this difference had the samples come from the same rather than different speakers: weak support for the prosecution. Thus the LR magnitude in this example is still not very big, even though the offender and suspect values are fairly similar and atypical.

Another common word in forensic samples of probably many varieties of English is *fuck* or *fucken*. Fig. 5 shows details from another acoustic–linguistic comparison between the F-pattern of the short open /ɐ/ vowel (often transcribed /ʌ/) in a set of seven *fucken*s recorded during a hold-up and three sets of *fucken*s intercepted from separate telephone calls involving the suspect. The F-pattern was sampled at 25% points of the duration of the nucleus. The vowels in the criminal sample sounded backer than those in the suspect samples, and this difference corresponds to the clear difference in relative position of F1 and F2. Table 1 gives the numerical data (means, standard deviations, number in sample) for the first three formant centre-frequencies measured at the mid-point of the vowel, both for offender sample, suspect samples and reference distribution. The reference distribution against which the differences between the samples were compared consists of formant data from a relatively large number of male Australian English speakers (Bernard, 1967). Two sets of reference distribution values are given, in Table 1, corresponding to the two alternative hypotheses entertained: the offender is a broad-speaking male other than the suspect (denoted by B); and the offender is someone other than the suspect with a non-cultivated

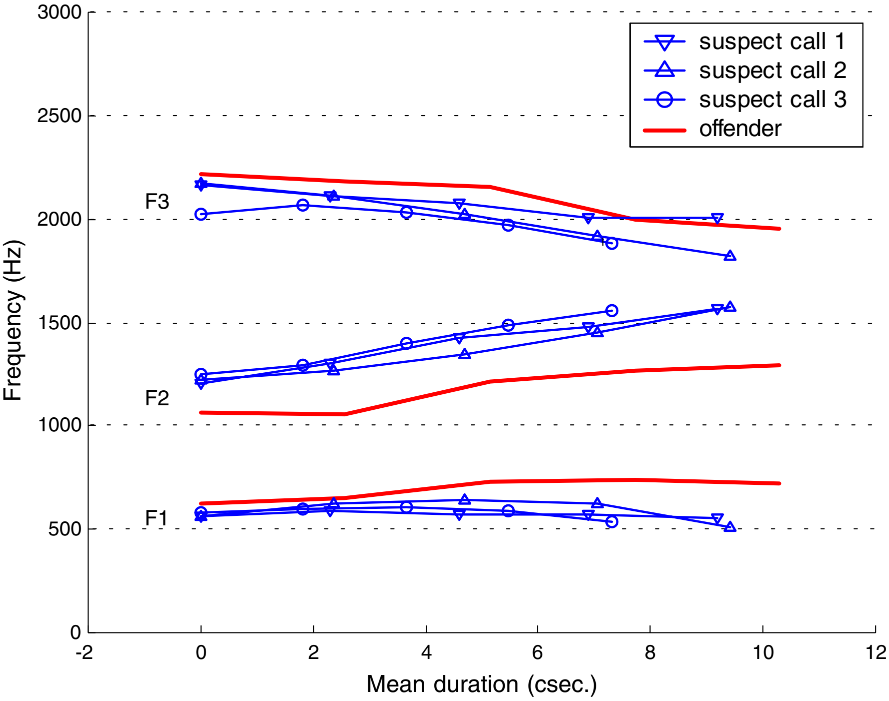

Fig. 5. Comparison between time course of mean F-pattern of /ɐ/ in offender *fucken* (thick line) and mean F-patterns of /ɐ/ in *fucken* from three intercepted suspect phone calls (thin lines).

Table 1
Data for LR comparison of mid-nucleus F-pattern in suspect (S) and offender (O) samples of /ɐ/ in *fucken*

		F1			F2			F3		
		x	sd	*n*	*x*	sd	*n*	*x*	sd	*n*
O		734	92.1	7	1215	99.8	6	2153	59.6	4
S	C1	574	28.0	3	1426	43.3	3	2072	24.5	3
	C2	621	38.4	5	1346	67.3	5	2021	97.2	5
	C3	611	57.1	14	1399	74.4	13	2029	159.0	11
R	B	737	69.4	56	1416	93.1	56	2526	146	56
	B + G	744	68.5	117	1414	84.4	117	2513	151.2	118

C1–C3 = suspect conversations 1–3. R = reference data for Broad (B) and combined Broad and General (B + G) Australian male /ɐ/ F-pattern. *x* = mean (Hz), sd = standard deviation (Hz), *n* = number in sample.

accent (denoted by B + G). (Australian accents are customarily classified on the basis of the quality of some vowels into three types, called Broad, General and Cultivated. In the case of the /ɐ/ vowel being tested, it can be seen that there is little difference between Broad and General values, and the results will therefore be very similar for both alternative hypotheses.)

Fig. 6 shows the mean F2 values involved against a reference distribution, modelled normally, of /ɐ/ F2 from 118 Broad and General Australian males. (A kernel density modelling was not used in this case, as its use in estimating LRs requires estimating within-speaker variance for the reference sample, which is problematic with the Bernard (1967) data, and in any case the distribution looks fairly normal. The reference distribution modelled with a Gaussian kernel density is shown in the insert to Fig. 6.[4])

It can be seen in Fig. 6 that the suspect's three mean F2 values are fairly typical, but that the offender's mean F2 is atypically low. It can also be seen that the difference between the suspect's means in conversations 1 and 2 is quite large. The variances involved differ a little, but as in the previous example, the mean within-speaker variation is generally about the same as the between-.

LRs were estimated for comparisons using each of the first three formants. A pooled-variance version of the LR formula at (3) was used, which assumes normality and equal variances (Rose, 2003, p. 184, 200). LRs were estimated – not only for the important offender-suspect comparison, but also for the within-suspect comparisons: any councel worth their salt would check how the known data were evaluated by the method. Quite apart from being a necessary part of the investigation, the demonstration of correct discrimination of known data can be lead as evidence in court and encourages confidence in results; incorrect discrimination of known data will, and should be, devastating under cross and demolish credibility.

Results for the LR comparisons with the three *fucken* /ɐ/ formants are in Table 2. This shows, for example, that when comparing the /ɐ/ F1 means in the suspect's two conversations C1 and C2, the

[4] Results for an attempt at a kernel-density estimate for these data were given in Rose (2004b), where it can be seen that they differ considerably in magnitude from those obtained with the less complicated model, although they agree in assessing the differences between the known suspect conversations as more likely assuming the same speaker, and differences between offender and suspect conversations as more likely assuming different speakers.

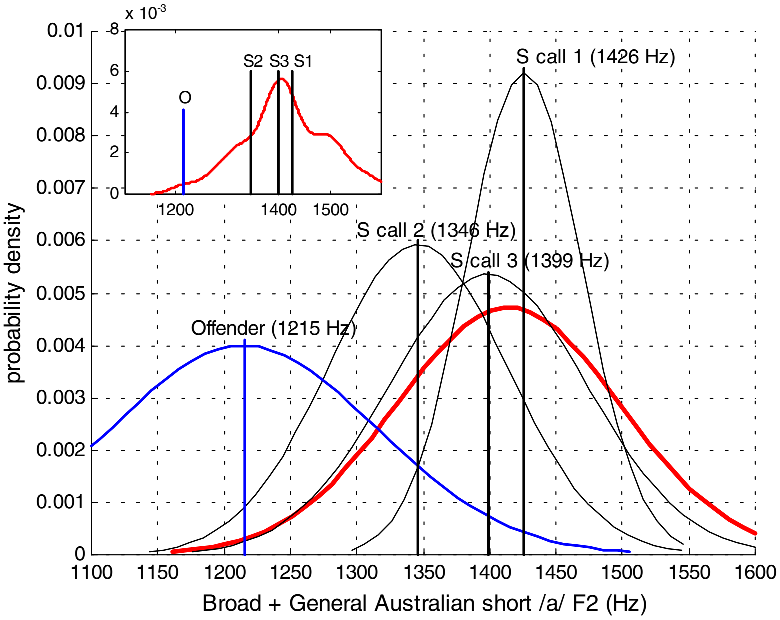

Fig. 6. Foresenic evaluation of an acoustic–linguistic feature (F2 target of /ɐ/ in *fucken*). Three suspect and one offender samples (thin lines) compared against a reference distribution from 118 speakers (thick line). Insert shows reference distribution modelled as Gaussian kernel density ($\lambda = 0.3$).

Table 2
Likelihood ratios for /ɐ/ F-pattern comparisons between suspect and offender *fucken* (S vs. O) and within-suspect *fucken*

Within-suspect	F1		F2		F3		Combined LR	
	B	B+G	B	B+G	B	B+G	B	B+G
C1 vs. C2	6.0 SS	7.4 SS	**1.9 DS**	**2.1 DS**	312 SS	176 SS	985 SS	620 SS
C1 vs. C3	14.4 SS	18.2 SS	1.7 SS	1.5 SS	204 SS	117 SS	4994 SS	3194 SS
C2 vs. C3	13.0 SS	11.7 SS	1.1 SS	1.1 SS	660 SS	350 SS	9438 SS	4505 SS
S vs. O	4.3 DS	3.7 DS	14.7 DS	15.5 DS	11.2 SS	6.8 SS	6 DS	8 DS

C1 = suspect conversation 1, etc. *n* SS/DS = *n* times more likely to observe difference between samples if from same speaker/different speaker. B, B+G = LRs for different alternative hypotheses (see text). Bold indicates LRs counter to known reality.

difference between their values would be about six times more likely were they from the same than different speakers, assuming an alternative hypothesis H_a that the offender was a Broad (B) speaker, and about seven times more likely, assuming the offender was a speaker from either the Broad or General (B + G) population. Since it is known that the data are in fact from the same speaker, this is an encouraging result. Note, however, that this is not the case with the F2 results for C1 vs C2, where the difference between the values is in fact marginally more typical for different speakers

(LRs = 1.9/2.1). This is partly a function of the fact that, as noted for Fig. 6 above, the F2 means for C1 and C2 are quite far apart, and the variances involved are relatively small. The fact that the LRs are still not big is largely because the difference between the means is still fairly typical.

When the values for all three formants in the suspect's speech are combined, in the right-most columns of Table 2, the differences are clearly considerably more likely assuming same-speaker provenance, and this is consistent with the known facts. (The combined LR is the product of the individual LRs assuming independent evidence; the DS (different speaker) LR values for F2 must be converted back to their original, reciprocal form.)

Having demonstrated that the approach gives the correct result with the known data, the questioned data can be addressed. In the comparison between the offender and suspect samples, the combined LRs of 6 (B) or 8 (B + G) indicate weak support for the defence hypothesis that they have come from different speakers (note again that the differences between the F3 values are more likely to have been observed assuming same-speaker provenance). The LR for this *fucken* /ɐ/ F-pattern feature is now available for combination with other LRs from the speech evidence.

It is essential to point out that, for several reasons, this is actually a very crude estimate indeed of the LR for this small piece of evidence. The reasons are as follows. Firstly, the samples have been compared with respect to F-pattern at only one point in the vowel – it is like a poor man's text-dependent speaker identification! (Comparison at other points is difficult because of lack of reference data.) Fig. 5 shows, however, that there are differences between suspect and offender's F-pattern throughout the formants' time course, so LRs taken at other points would probably also show greater support for the defence hypothesis.

Secondly, because the suspect data were obtained from phone intercepts, it could be objected that their F1 should not have been included due to the well-known potential band-pass effect which tends to shift F1 estimates up, especially for high and mid vowels (see Fig. 3). However, it can be seen in Fig. 5 that the suspect's F1 is actually lower than the offender's, so if there has been any band-pass shifting, it would have brought the suspect's F1 nearer the offender's, and been in favour of the prosecution.

Thirdly, the reference data are not totally comparable to the forensic data: the reference data are for stressed /ɐ/ vowels before a final alveolar consonant as in *hut*, whereas the /ɐ/ vowel in the samples occurs before a velar.

Next must be reiterated the shortcomings – mentioned in section 5 above – of the LR formula. This can best be seen from a comparison with results obtained from the attempt at a kernel-density estimate mentioned in footnote 4. Although both approaches agree in their predictions, the kernel density estimate would have it that the differences between the offender and suspect are ca. 770 times more likely assuming they have come from different speakers, compared to the factors of 6/8 for the formula assuming normality! Although this discrepancy is probably due more to problems in estimation of the between-speaker variance than the formula itself, it does show how dependant our figures are on the modelling, and that a FSR case should never rely on comparison of a single feature, or even a few features alone.

Finally, in implementing the "Idiot's Bayes" approach of simply taking the product of the LRs to estimate a combined LR, no account has been taken of possible correlations between different formant measurements.

All these shortcomings make it even more important to be able to show that the correct discrimination is obtained with the known comparisons.

8.2. An acoustic–non-linguistic feature

An acoustic–non-linguistic feature often used in forensic comparison is long term average F0 (LTF0). Although it is possible to consider LTF0 as a linguistic feature because it is known to characterise different languages, it is probably best regarded as non-linguistic because it strongly reflects both *Intrinsic Indexical* features like length and mass of the cords, and state of health, as well as non-linguistic aspects of *Communicative Intent* like *Affect* and *Self-presentation*. (The italicised terms are part of an explicit model for the information content in a voice (Nolan, 1983, 2002, Chapter 10) – a third conceptual framework which, together with Bayes' theorem and Linguistics, underlies non-automatic TFSR).

Fig. 7 represents a forensic comparison between suspect and offender in mean LTF0, again using kernel density estimation. The language is Cantonese. The suspect's LTF0 is the mean of 14 phone calls in which he acknowledged he participated; the offender's value is from one phone call adjudged long enough to provide a good estimate of his LTF0 (Rose, 1991). The reference distribution is from means of 17 Cantonese males speaking over the phone (Rose, 2003, pp. 4110–4111). The 2.3 Hz difference between the offender and suspect LTF0 is extremely small – it represents only about 2% of a male Cantonese speaker's typical range (2 * LTF0sd) (Rose, 2000). It is also easily of a magnitude that could be caused by a change in the settings for automatic F0 extraction. However, the values also lie near the reference distribution mode and are thus fairly typical, and once again there is little difference between the within- and between-speaker variances. According to the kernel density LR formula at (4), one would only

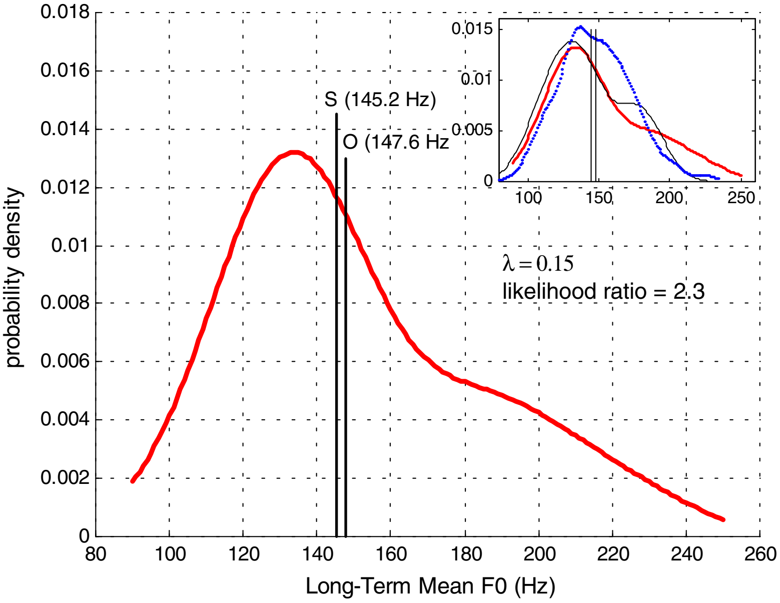

Fig. 7. Mean suspect and offender LTF0 samples compared against a GKD reference distribution of Cantonese LTF0 from 17 males. Insert shows GKD distributions of suspect's LTF0 means (14 phone conversations, solid line) and the F0 distribution in the single offender call (dotted line).

be about twice as likely (LR = 2.3) to observe this difference were the samples from the same speaker – on its own, nearly useless as evidence. This is a good example of why similarity between samples is only half the story in forensic comparison.

8.3. Examples of auditory features

There is effectively a limitless number of potential auditory features that can be used in the forensic comparison of speech samples. Table 3 contains some typical examples of differences observed between offender and suspect samples in a case involving Chinese (Rose, 2003, p. 4063–4068). It is worth noting that the voice in both samples sounded very similar in non-linguistic features like overall pitch and phonation type – similarities that one would perhaps be more likely to observe were they from the same speaker.

The first example in Table 3 is of differential placement of the time adverb xiān *first*: pre-verbal in suspect sample; post-verbal in offender. Examples 2 and 3 are of a consistent difference between a word-initial retroflex fricative [ʂ] (suspect) and an alveolar fricative [s] (offender). This reflects a more general phonological situation where the offender's sample lacks a whole set of phonemic contrasts between retroflex and alveolar syllable-initial consonants that is present in the suspect sample. Example 4 shows a correspondence between syllable-initial alveolar nasal [n] (suspect) and lateral [l] (offender).

It is difficult to convey the flavour of these differences. They are loosely analogous to a situation with British English where the voice in one sample has the two "th" sounds [ð] and [θ] (as in *this* and *thing*), and the voice in the other sample does not. The speaker in the second sample would make no difference between words like *that* and *vat*, saying them both as *vat*; and between words like *thought* and *fought* saying them both as *fought*.

The features in the offender sample are in fact typical of a Cantonese speaker speaking Standard Chinese (Cantonese does not have retroflex consonants; typically lacks [n] syllable-initially; and puts time adverbs after the verb). Part of the background information of the case is that the suspect was born and grew up in Peking, where they do have retroflexes and syllable-initial [n], and prepose time adverbs (and his speech reflects that).

Table 3
Example of auditory-linguistic comparison of forensic voice samples in Putonghua (Standard Chinese)

	Suspect's samples		Offender's samples	
	[utterance]	Transcription Chinese/(Pinyin) and *gloss*	[utterance]	Transcription Chinese/(Pinyin) and *gloss*
1	ni ɕɛn tɕ^{h}y pa	你先去吧 (nǐ xiān qù ba) *better you go first*	tɐŋ iɕa ɕɛn a	等一下先啊 (děng yíxià xiān a) *wait a bit first*
2	ʂwɔ	說 (shuō) *speak*	swɔ	說 (shuō) *speak*
3	ʂʐwu	十五 (shíwǔ) *fifteen*	ʂzxɔʊ	時候 (shíhou) *time*
4	na	那 (nà) *In that case*	lali	哪里 (nǎlǐ) *where?*

In order to evaluate the differences between the two samples, one must ask what the probabilities are of observing them assuming they were spoken by the same speaker; and assuming they were not. It is difficult, though not impossible, to conceive of a situation where the same speaker might show these features in two different conversations. Speakers are known to converge and diverge as a normal part of conversational interaction: it is one way of signalling in- or out-group membership. Thus these differences might conceivably arise as the result of either convergence or divergence. The former hypothesis is not possible, as a linguistic analysis of the speech of the offender's interlocutor shows that he actually distinguishes retroflexes, and n from l. This leaves the divergence hypothesis, for which there is little evidence, as the speakers sound as if they are getting on fine.

It is clear, then, that these auditory-linguistic differences would be far more likely under the assumption that the samples had come from different speakers. This would certainly be at least enough to balance the "same-speaker" LR that might come from consideration of the abovementioned similarity between the samples in voice quality features, and probably constitute strong support for the defence.

9. Evidentiary standards and testing

The by now well-known USA Supreme Court *Daubert* rulings on admissibility of scientific evidence (Daubert, 1993) include, as one criterion, whether the theory or technique can be, and has been, tested (Black et al., 1994, pp. 750–776). In Federal and State Australian courts the practice notes requiring reliability, replicability and transparency on the part of expert testimony are *de facto* adoptions of *Daubert*. It is a natural question, therefore, to ask to what extent the Likelihood Ratio approach to Forensic Speaker Recognition described in this paper has been tested. (There are other important questions to be asked concerning the relationship between the *Daubert* criteria and the Bayesian evaluation of scientific forensic evidence – see Robertson and Vignaux, 1995, 205ff.)

The idea of testing a theorem is not coherent since it does not possess the property of being wrong, and its truth is guaranteed (Robertson and Vignaux, 1995, p. 17; Royall, 2000, p. 760). Rather, it is that part of the analytical approach which has to do with the extraction and quantification of the differences between the samples that can be tested. Given that the Likelihood Ratio is predicted to be greater than unity for same-subject data, but less than one for different-subjects, it can be used as a discriminant distance around the appropriate threshold (1 or 0), and the evidence consisting of known same-speaker and different-speaker pairs tested to see to what extent they are correctly resolved – a relatively straightforward discrimination between same-speaker pairs and different-speaker pairs.

There has already been experimentation of this kind – and not only on speech. For example Evett et al. showed in 1993 that repeat DNA samples from 152 subjects were, as predicted, resolved with Likelihood Ratios greater than 1, whereas ca. 1.2 million pairs of DNA samples from different subjects were, again as predicted, associated with Likelihood Ratios of less than 1 in the vast majority of cases (only eight in a million comparisons of DNA from different subjects yielded a LR greater than 1). Similar, although not so large-scale, LR-based discriminant experiments have been performed on other forensically common trace material, for example elemental ratios in glass fragments (Brown, 1996; Aitken and Lucy, 2004; Aitken et al., in press). These experiments do not boast quite so spectacular results. Aitken et al. (in press) for example found a "disappointingly low" proportion of true positives, with a correspondingly high false negative rate. This highlights another problem with testing the LR-based approach, namely that it is not always easy to separate out the natural discriminability of the data from the adequacy of the discriminant method used: there probably is not as much individual-identifying content in elemental ratios of glass as in DNA, but perhaps the methods used were not adequately tapping the individual-identifying content in the glass.

Likelihood-ratio based discriminant experiments on speech have been considerably more successful, with results clearly supporting the hypothesis that same-speaker data can be well discriminated from different-speaker data using a Likelihood Ratio (see, e.g., Gonzalez-Rodriguez et al. in this volume). This success is to be expected from the long line of ever diminishing EERs in automatic verification experiments – for example the NIST evaluations – which has shown that same-speaker pairs can be discriminated from different-speaker pairs with considerable reliability, under fairly tough, even forensically authentic, conditions. For example Leeuwen and Bouten (2004, p. 75, 76–77) report a lowest EER of 12.1% in experiments with real forensic data involving 40 speakers, 521 target- and 9676 non-target-trials, and Gonzalez-Rodriguez et al. (this volume) demonstrate high discriminant power even in cases where there is only one questioned and one suspect recording available for comparison.

Most of the LR-based discriminant experiments used to investigate the approach have been carried out with automatic methods, using non-linguistic evidence. For example Meuwly and Drygajlo (2001, p. 149) tested eight Swiss French speaking males against themselves and a reference distribution of 1000 males in an experiment involving 48 same-speaker and 8000 different-speaker trials. They found (Meuwly and Drygajlo, 2001, p. 150) that ca. 86% of the same-speaker comparisons had LRs bigger than unity and ca. 88% of different-speaker comparisons LRs less than unity.

More recently, Leeuwen and Bouten (2004, p. 77, 81–82) included Likelihood Ratio-based discrimination in their evaluation of the performance of automatic approaches on real forensic data. They tested 10 Dutch speakers against some of the other non-targets in their corpus, making 287 same-speaker comparisons, and 2353 different-speaker (i.e., non-target) comparisons (it is not clear how many non-targets were used.) They present a figure (Leeuwen and Bouten, 2004, p. 82) with results for three of the systems evaluated that they say represent a wide range of performance. The best of these three systems resolves ca. 96% of same-speaker comparisons with LRs bigger than unity and ca. 96% of different-speaker comparisons with LRs less than unity. The worst system shown has ca. 78% of different-speaker comparisons with $LR < 1$ and ca. 68% of same-speaker comparisons with $LR > 1$. They point out (Leeuwen and Bouten, 2004, p. 82) that maximal separation between the discriminant performance in same- and different-speaker pairs occurs at threshold, which is where theory predicts it to be.

Several Likelihood Ratio-based experiments have also been carried out with linguistic evidence. All of them have used an analytical LR formula, rather than an empirical approach. Kinoshita (2001, 2002) provided the first demonstration of the method using formant centre-frequencies. She tested 90 same-speaker pairs and 180 different-speaker pairs from ten male Japanese speakers, and found 90% of same-speaker trials with $LR > 1$, and 97% of different-speaker pairs with $LR < 1$. She was able to obtain these results with just six formant centre-frequency measurements.

Although she used non-contemporaneous natural speech, it was not of telephone quality. This may have influenced the results, but it is unlikely, given that she used formants that were unlikely to be adversely affected by the telephone pass-band. In a similar small-scale experiment using the F-pattern of the five Australian English long vowel phonemes, Alderman, 2004a, p. 181 tested 11 same-speaker and 220 different-speaker pairs from 11 male speakers' non-contemporaneous speech (including one pair of identical twins). When evaluated against a reference distribution of over 100 speakers he obtained only ca. 70% of same-speaker pairs with a LR > 1, but ca. 99% of different speakers with LR < 1. Again, this result was obtained with just five F2 measurements, and F1 in /a/ and /ɜ/.

A slightly larger-scale experiment on forensically realistic data was recently carried out with non-contemporaneous phone recordings from 60 Japanese males, involving 240 same-speaker and 28,320 different-speaker trials (Rose et al., 2003). Only three phonetic segments were used – a vowel [ɔ], a voiceless fricative [ɕ] and a nasal [ŋ], and the approach was therefore characterised as segmental-acoustic. LRs were estimated for two kinds of analysis commonly found in TFSI – F-pattern and cepstrum. For the cepstral LR, only 38% of same-speaker pairs had likelihood ratios greater than 1, but notably only 0.04% of different speakers had LRs greater than 1. As stressed in Gonzalez-Rodriguez et al. (2004, p. 84, *et pass*) this very low possibility of error may be a desirable property of LR-based approaches in the real forensic world, where one wants to avoid incriminating the innocent. On the other hand, the inevitably high missed-hit rate is still a worry.

Auditory-linguistic and non-linguistic features in the word *okay* in Australian English have also been tested in this way, and found to yield low but useable LRs (Elliott, 2002). The features, which are categorical, were observed in the spontaneous speech of ten young Australian males in two separate conversations separated by at least several months, and include: palatalisation of /k/ to fronted velar [c] – i.e., "*o*[cʰ]*ay*"; instead of "*o*[kʰ]*ay*"; frication of /k/ to velar fricative [x] – i.e., "*o*[x]*ay*"; voicing of /k/ to [g] – i.e., "*o*[g]*ay*"; realisation of first diphthong /oʊ/ as "[ɛɪ]-*kay*"; nasalisation of diphthongs; use of creaky voice. The Likelihood Ratios associated with these features are such that a certain incidence of, say, palatalisation in both offender and suspect samples would be nine times more likely if they were from the same speaker. Since the same average Likelihood Ratio was found for the creaky voice feature, and the two features appear to be independent, two samples both with creaky palatalised *okay*s – i.e., [aʉ̯̰cʰɛɪ̯̰] would be 81 times more likely assuming same speaker.

A Likelihood Ratio value is a prediction based on statistical inference, and such predictions can be wrong even though the inference itself is correct. The error rates quoted above show that it is possible in a comparison of same-speaker samples to get a LR smaller than one, and *vice versa*. The probability of getting misleading statistical evidence in connection with Likelihood Ratios was addressed analytically in important papers by Lindley, 1977, pp. 210–211 and Royall (2000). Error probability in automatic FSR is usually assessed empirically, however, and conventionally illustrated by means of so-called reliability, or Tippett plots. These are ogives (cumulative distributions) of Likelihood Ratios from same-subject and different-subject trials, usually with the same-subject Likelihood Ratios plotted inversely. There are many examples in Gonzalez-Rodriguez et al. (this volume). The graphs thus show for what proportion of same- or different-speaker trials one observes a Likelihood Ratio bigger than a given abscissa LR value. It is claimed (Drygajlo et al., 2003) that such graphs can be used to indicate to a court how reliable a Likelihood

Ratio value quoted in a particular case is, but they probably reflect the reliability of the system more. It should not be forgotten that such calibration experiments, and the Tippett plots summarising their results, tap *overall* behaviour and cannot say anything about the accuracy of a LR *in a particular case* (Rose, 2002, pp. 318–325; 2003, p. 2057). Up to now, proposals for evaluating once-off probabilities of this kind have not been successful (Royall, 2000, p. 760). Tippet plots, although popular, have also been criticised for their sensitivity to details in their implementation, particularly the "...statistically dubious re-use of speech samples in different roles" (Leeuwen and Bouten, 2004, p. 82). Nevertheless, Tippett plots remain valuable for reminding us that our derived Likelihood Ratio values are still, and inevitably, associated with error.

Since the focus of this section has involved one aspect of the relationship between Science and the Law (legal admissibility of scientific evidence), it is worth remembering that there are many areas of tension, even conflict, between them. There is, firstly, the uneasy relationship – alluded to in the previous paragraph – between the surely reasonable requirement of testability under *Daubert* and the essentially unique circumstances of each case. The different legal construal of the null hypothesis (the speech samples must be assumed to have come from different speakers until prosecution convinces otherwise) is another example. The very application of probability is yet another (Dawid, 2005, pp. 7–8): sooner or later, for example, the scientific probability estimate has to collapse into a binary judgement. Whom the court appoints or recognises as expert may not be considered a *bona-fide* expert by the general forensic speaker recognition community.

In procedural and interpretive matters also, descriptive Science and prescriptive Law do not always see eye to eye, and ultimately the Law – which is "what judges say it is" (Good, 2001, preface) – must of course prevail. A recent UK ruling in O'Doherty (2002) for example, is that in TFSR cases acoustic evidence must be brought – a good thing, given that it also represents the consensus of opinion among practicing forensic phoneticians that both auditory and acoustic evidence must be used (Rose, 2002, p. 35). The judgment on inadmissibility of evidence based on Bayes' theorem quoted in section 2 above is not such a good thing: it has the potential to introduce all sorts of problems for LR-based TFSR evidence unless the judiciary can be convinced that it is not Bayesian in the formal sense.

10. Summary

This paper has discussed some important aspects of Technical Forensic Speaker Identification, focusing on both the necessary logical framework for evaluation of forensic speaker identification evidence, and how non-automatic methods, using true higher-level linguistic knowledge, can be of forensic use. The main message, I think, given the excellent performance of automated systems, is nevertheless that not all evidence is being exploited in estimating Likelihood Ratios. It is clear that the traditional approaches, which lack the globality of the automatic approaches, are not extracting all information relevant to the estimation of a LR. It is equally clear that automatic methods, which by definition do not take true higher level linguistic information into account, may also be missing information, since it has been shown that this higher level information can furnish on its own strong LRs in support of either defence or prosecution. If it is conceded that a LR be estimated for all possible information in a FSR case, then ideally both linguistic and automatic approaches should be combined. Some fruitful collaboration is in order, envisaged and encouraged by the addition of *acoustics* in the recent name change from the *International Association for Forensic Phonetics* to the *International Association for Forensic Phonetics and Acoustics.*

Perhaps, also, a final caveat is in order, given the display of discriminant power evidenced above. Does it need to be emphasised that, despite the enormous advances in technology, the intimidatingly complex statistical modelling, and impressive discrimination scores, Forensic Speaker Recognition is still a complex undertaking, with serious consequences, that should only be done by those who are aware of its limitations? Most decidedly so.

Acknowledgements

This is an expanded version of the paper (Rose, 2004a) accompanying my keynote presentation on Forensic Speaker Recognition at the *Odyssey Speaker Recognition* workshop in Toledo 2004. I thank Dr. James Robertson AO, Head of the Forensic Service of the Australian Federal Police, and Hugh Selby, Reader in Law at the Australian National University, for useful discussions about evidence. I also want to acknowledge the contribution of my co-presenter Dr. Didier Meuwly (late of the Forensic Science Service and now at the Netherlands Forensic Institute), especially for generously offering his DNA data represented in Fig. 1. I also thank Professor Dennis Lindley, Dr. Daniel Ramos-Castro, Dr. Peter French, Dr. Ann Kumar and Tony Alderman for suggesting references/obtaining papers for me. Special thanks to all my seven reviewers for giving their time to carefully read and make knowledgeable comments which have helped make this a better paper. The views expressed in this paper are of course my own.

References

Aitken, C.G.G., 1995. Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley & Sons, Chichester.

Aitken, C.G.G., Lucy, D., 2004. Evaluation of trace evidence in the form of multivariate data. Applied Statistics 53/4, 109–122.

Aitken, C.G.G., Lucy, D., Zadora, G., Curran, J.M., (in press). Evaluation of transfer evidence for three-level multivariate data with the use of graphical models. Computational Statistics and Data Analysis.

Aitken, C.G.G., Stoney, D.A., 1991. The Use of Statistics in Forensic Science. Ellis Horwood, Chichester.

Aitken, C.G.G., Taroni, F., 2004. Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, Chichester (second ed. of Aitken 1995).

Alderman, T., 2004. The use of Australian-English vowel formant data sets in forensic speaker identification. In: Cassidy, S. (Ed.), Proceedings of the 10th Australian International Conference on Speech Science and Technology (PANZE workshop), pp. 177–182.

Alderman, T., 2004. The Bernard data set as a reference distribution for Bayesian Likelihood-ratio-based forensic speaker identification using formants. In: Cassidy, S. (Ed.), Proceedings of the 10th Australian International Conference on Speech Science and Technology, pp. 510–515.

Balding, D., 2005. A question of identity. Significance (March), 20–23.

Bernard, J.R.L., 1967. Some measurements of some sounds of Australian English. Unpublished Ph.D. Thesis, University of Sydney.

Bernado, José M., 2001. Bayesian statistics. In: Encyclopedia of Life Support Systems. UNESCO.

Black, B., Ayala, F.J., Saffran-Brinks, C., 1994. Science and the law in the wake of *Daubert*: a new search for scientific knowledge. Texas Law Review 72/4, 715–802.

Boë, L.-J., 2000. Forensic voice identification in France. Speech Communication 31, 205–224.

Bolt, R.H., Cooper, F.S., David Jr, E.E., Denes, P.B., Pickett, J.M., Stevens, K.N., 1970. Speaker identification by speech spectrograms: a scientists' view of its reliability for legal purposes. JASA 47/2, 597–612.

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J., Reynolds, D.A., Magrin-Chagnolleau, I., 2003. Person authentication by voice: a need for caution. In: Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH).

Broeders, A.P.A., 1999. Some observations on the use of probability scales in forensic identification. Forensic Linguistics 6/2, 228–241.

Broeders, A.P.A., 2001. Forensic speech and audio analysis forensic linguistics – 1998–2001: a review. Paper at the 13th INTERPOL Forensic Science Symposium.

Broeders, A.P.A., 2004. Presentation. Workshop on Evidence & Identity. Joseph Bell Centre for Forensic Statistics and Legal Reasoning. University of Edinburgh.

Brown, K., 1996. Evidential value of elemental analysis of glass fragments. Unpublished first class Honours Thesis, University of Edinburgh.

Byrne, C., Foulkes, P., 2004. The 'mobile phone effect' on vowel formants. Speech Language and the Law 11/1, 83–102.

Champod, C., Evett, I.W., 2000. Commentary on Broeders (1999). Forensic Linguistics 7/2, 238–243.

Daubert, 1993. Daubert vs Merrell Dow Pharmaceuticals, Inc. 113 S Ct 2786.

Dawid, P., 2005. Statistics on trial. Significance (March), 6–8.

Drygajlo, A., Meuwly, D., Alexander, A., 2003. Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. In: Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH).

Elliott, J., 2001. Auditory and F-pattern variations in Australian *okay*: a forensic–phonetic investigation. Acoustics Australia 29/1, 37–41.

Elliott, J., 2002. Okay, what are the odds? Unpublished M.Phil. Thesis. Australian National University.

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

Evett, I.W., Scrange, J., Pinchin, R., 1993. An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. American Journal of Human Genetics 52, 498–505.

French, P., Harrison, P., 2004. The *who wants to be a millionaire?* fraud trial. Speech Language and the Law 11/1, 131–145.

Gigerenzer, G., 2002. Reckoning with Risk. Allen Lane The Penguin Press, London.

Gigerenzer, G., Hoffrage, U., 1995. How to improve bayesian reasoning without instruction: frequency formats. Psychological Review 102/4, 684–704.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance. CUP, Cambridge.

Gonzalez-Rodriguez, J., Ortega-Garcia, J., Sanchez-Bote, J.-L., 2002. Forensic Identification Reporting Using Automatic Biometric Systems. In: Zhang, D., Sarukkai, S.R. (Eds.), Biometric Solutions for Authentication in an E-world, Kluwer International Series in Engineering and Computer Science, 697. Kluwer, Dordrecht, pp. 169–186.

Gonzalez-Rodriguez, J., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J., 2004. On robust estimation of likelihood ratios: the ATVS-UPM system at 2003 NFI/TNO forensic evaluation. In: Ortega-Garcia, J. et al. (Eds.), pp. 83–90.

Good, P., 2001. Applying Statistics in the Courtroom – A New Approach for Attorneys and Expert Witnesses. Chapman & Hall/CRC, London.

Hand, D.J., Yu, Keming, 2001. Idiot's Bayes — not so stupid after all? International Statistical Review 69/3, 385–398.

Hodgson, D., 2002. A lawyer looks at Bayes' theorem. The Australian Law Journal 76, 109–118.

Kinoshita, Y., 2001. Testing realistic forensic speaker identification in Japanese: A Likelihood ratio based approach using formants. Unpublished Ph.D. Thesis, The Australian National University.

Kinoshita, Y., 2002. Use of likelihood ratio and Bayesian approach in forensic speaker identification. In: Bow, C. (Ed.), Proceedings of the 9th Australian International Conference Speech Science and Technology. Australian Speech Science & Technology Association, Melbourne, pp. 297–302.

Köller, Nissen, Rieß, 2004. Sadorf. Probabilistische Schlußfolgerungen in Schriftgutachten. Luchterhand: Polizei und Forschung 25.

Künzel, H.J., 2001. Beware the telephone effect: the influence of transmission on the measurement of formant frequencies. Forensic Linguistics 8/1, 80–99.

Künzel, H.J., González-Rodríguez, J., 2003. Combining automatic and phonetic–acoustic speaker recognition techniques for forensic applications. In: Proceedings of the 15th ICPhS 2003, pp. 1619–1622.

Ladefoged, P., 2004. The law is not science: the validity of voice identification. JASA Echoes 14/2, 14–15.

Leeuwen, D.A., Bouten, J.S., 2004. Results of the 2003 NFI-TNO forensic speaker recognition evaluation. In: Ortega-García, J. et al. (Eds.), pp. 81–82.

Lewis, S.R., 1984. Philosophy of Speaker Identification. Proc Institute of Acoustics 6/1, 69–77.

Lindley, D.V., 1977. A problem in forensic science. Biometrika 64/2, 207–213.

Lindley, D.V., 1991. Probability. In: Aitken, C.G.G., Stoney, D.A. (Eds.), pp. 27–50.

McDougall, K., 2004. Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. Speech Language and the Law 11/1, 103–130.

Meuwly, D., 2001. Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. Unpublished Ph.D. thesis, University of Lausanne.

Meuwly, D., 2004a. Forensic speaker recognition – an evidence Odyssey: summary. In: Ortega-García, J. et al. (Eds.), pp. 11–12.

Meuwly, D., 2004b. Keynote presentation and associated overheads, Forensic Speaker Recognition Workshop, Speaker Odyssey '04.

Meuwly, D., 2005. 'Biometrics in forensic science: fingerprint, voice, DNA', invited presentation, *Identech*.

Meuwly, D., Drygajlo, A., 2001. Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM). In: Proceedings of the 2001 Speaker Odyssey – Speaker Recognition Workshop, pp. 145–150.

Meuwly, D., El-Maliki, M., Drygajlo, A., 1998. Forensic speaker recognition using Gaussian mixture models and a Bayesian framework. COST-250 Workshop, Ankara.

Nakasone, H., Beck, S.D., 2001. Forensic automatic speaker recognition. In: Proceedings of the 2001 Speaker Odyssey Speaker Recognition Workshop.

Nolan, F., 1983. The Phonetic Bases of Speaker Recognition. CUP, Cambridge.

Nolan, F., 1990. The limitations of auditory-phonetic speaker identification. In: Kniffka, H. (Ed.), Texte zur Theorie und Praxis forensischer Linguistik. Max Niemayer Verlag, Tübingen.

Nolan, F., 2003. A recent voice parade. Speech Language and the Law 10/2, 277–291.

Nolan, F., Oh, T., 1996. Identical twins, different voices. Forensic Linguistics 3/1, 39–49.

Ortega-García, J., González-Rodríguez, J., Bimbot, F., Bonastre, J.-F., Campbell, J, Magrin-Chagnolleau, I., Mason, J., Peres, R., Reynolds, D., (Eds.), 2004. Proceedings of the Odyssey-04, The Speaker and Language Recognition Workshop.

Pinker, S., 1997. How the Mind Works. Penguin, London.

O'Doherty, 2002. R v O'Doherty. Her Majesty's Court of Appeal in Northern Ireland. [2002] NICA 20. Available from: <http://www.worldlii.org/nie/cases/NICA/2002/20.html>.

Doheny, 1996. R v Doheny. Court of Appeal Criminal Division. No. 95/5297/Y2.

Rish, I., 2001. An empirical study of the naive Bayes classifier. In: 17th International Joint Conference on Artificial Intelligence, Empirical Methods workshop, 2001, pp. 41–46.

Robertson, B., Buckelton, J., Dawid, P., 2005. Round-table debate on The Bayesian Approach to Evidence Evaluation. Human Identity Symposium HumiD. Available from: <http://www.e-symposium.com/humid/archive.php>.

Robertson, B., Vignaux, G.A., 1995. Interpreting Evidence. Wiley, Chichester.

Rose, P., 1991. How effective are long-term mean and standard deviation as normalisation parameters for tonal fundamental frequency? Speech Communication 10, 224–229.

Rose, P., 1997. Identifying criminals by their voice – the emerging applied discipline of forensic phonetics. Australian Language Matters 5/2, 6–7.

Rose, P., 1999. Differences and distinguishability in the acoustic characteristics of *hello* in voices of similar-sounding speakers: a forensic–phonetic investigation. Australian Review of Applied Linguistics 22/1, 1–42.

Rose, P., 2000. Hong Kong Cantonese citation tone acoustics – a linguistic–tonetic study. In: Barlow, M. (Ed.), Proceedings of the 8th Australian International Conference on Speech Science and Technology. Australian Speech Science & Technology Association, Canberra, pp. 198–203.

Rose, P., 2002. Forensic Speaker Identification. Taylor and Francis, London and New York.

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

Rose, P., 2004a. Technical forensic speaker identification from a Bayesian linguist's perspective. In: Ortega-García et al. (2004), pp. 3–10.

Rose, P., 2004b. Overheads accompanying keynote presentation, Forensic Speaker Recognition Workshop, Speaker Odyssey '04.

Rose, P., Lucy, D., Osanai, T., 2004. Linguistic–acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchial random effects model: a "non-idiot's Bayes" approach. In: Cassidy, S. (Ed.), Proceedings of the 10th Australian International Conference on Speech Science and Technology. Australian Speech Science and Technology Association, Sydney, pp. 492–497.

Rose, P., Osanai, T., Kinoshita, Y., 2003. Strength of forensic speaker identification evidence – multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold. Speech Language and the Law 10/2, 179–202.

Rose, P., Simmons, A., 1996. F-pattern variability in disguise and over the telephone: comparisons for forensic speaker identification. In: McCormack, P., Russell, A. (Eds.), Proceedings of the 6th Australian International Conference on Speech Science and Technology. Australian Speech Science and Technology Association, Canberra, pp. 121–126.

Royall, R., 2000. On the probability of observing misleading statistical evidence. Journal of the American Statistical Association 95/451, 760–768.

Sprent, P., 1977. Statistics in Action. Penguin, Harmondsworth.

Tanner, D.C., Tanner, M.E., 2004. Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection. Lawyers & Judges Publishing Company, Inc, Tucson, AZ.

Yarmey, D., 2004. Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalisations. Speech Language and the Law 11/2, 267–277.