Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud

Phil Rose

Abstract
The first use of likelihood ratios for the evaluation of forensic voice comparison evidence in a real trial in Australia is documented. Important steps in the process of estimating the strength of evidence – from acoustic-phonetic features in the utterances ‘yes’ and ‘not too bad’ – are described and explained. These comprise the nature and currency of likelihood ratios, choice of features, graphical demonstration of similarity and typicality of suspect and offender speech samples, and their multivariate likelihood ratio evaluation against a reference sample. Improvements in likelihood ratio estimation since the case are exemplified; in particular how to demonstrate that the approach actually works. The reception of the likelihood ratio-based evidence during the trial is used to highlight some problems in its rational evaluation by a court.

KEYWORDS FORENSIC VOICE COMPARISON, TELEPHONE FRAUD, LIKELIHOOD RATIO, F-PATTERN, FUNDAMENTAL FREQUENCY
How to make, or lose, $150 million in one phone call

On Christmas Eve 2003 a fraudulent fax was sent to the investment bank JP Morgan Chase in Australia requesting the transfer of $150 million to accounts in Switzerland, Greece and Hong Kong. About 10 minutes before the closing of business, the bank received a phone call from a Craig Slater, asking for a callback on the fax. This is a procedure confirming the details of the fax and verifying that the transfer could go ahead. The phone call started with the following exchanges (E = JP Morgan employee, F = fraudster):

E  J.P. Morgan Greg speaking
F  yeah hello Greg this is Craig Slater here mate
E  oh g’day how are you
F  not too bad I’ve been having a bit of trouble here we
   erh I haven’t been able to get onto anyone else on
   the other lines …
E  yep
F  is it would it be possible to you to do a call back
E  erh just a second I’ll just go check the fax
F  O. … O.K.

The bank employee then reads out from the fax the amounts to be transferred, and Slater tersely confirms them, for example:

E  erh and we’re going to pay Hong Kong dollars one one
   eight six seven eight five four three spot two nine
   [$118,678,543.29] to HSBC  erh Hong Kong
F  correct
   …
E  Hong Kong I think Hong Kong Power Limited six three
   six double oh three oh five five double oh one
   [$636,003,055,001]
F  yes

The call then ends with the appropriate season’s greetings:

E  is that correct
F  that’s correct
E  OK then
F  than .. thank you very much
E  have a good Christmas
F  you have a good Christmas too.  bye
E  OK bye
The Australian Commonwealth Superannuation Scheme account administered by the bank was now short by $150 million.

**Forensic voice comparison and the likelihood ratio**

The telephone fraud just described presents a typical scenario for forensic voice comparison (FVC). There was a recording of the unknown offender’s voice committing the fraud. The police had identified a suspect and were able to provide known recordings of their voice for comparison with that of the offender. The aim of the FVC was to help the trier-of-fact (in the case described here, a 12-person jury) decide whether the suspect had said the incriminating speech. Perhaps atypical is that, since most of the telephone call actually constituted the fraud, the prosecution and defence hypotheses were not only at the level of source and activity but of offence (Lucy 2005: 118 ff.).

Currently FVC, as reflected in research and practice, can be divided into two types depending on how the expert’s help is construed. In the first type, the expert is expected, and considers it their aim, to say how likely it is, given the evidence, that the suspect said the incriminating speech. Here is a typical conclusion from case-work (many similar examples can easily be found):

> the close matching of all respective parameters leads me to believe that there is strong evidence that the male speaker in the telephone calls and in the control recording are one and the same person.

In the second type of FVC, the expert’s aim is seen as restricted to estimating the strength of the speech evidence with a likelihood ratio (LR). This means estimating how much more likely the difference between the suspect and offender speech samples is, assuming the offender sample has come from the suspect, rather than from another randomly chosen speaker in the relevant population. Here is a LR-based conclusion in FVC:

> I estimate you would be about one thousand times more likely to observe the differences between the suspect and offender speech samples if the suspect, rather than someone else, had said the offender speech sample.

For some time now, it is the second type of help – with a LR – that has been theoretically recognised as the correct logical framework for the evaluation of forensic evidence. The LR is endorsed in all major textbooks on forensic statistics, e.g. Aitken and Taroni (2004), Balding (2005), Lucy (2005), as well as in the guide for judges, lawyers, forensic scientists and expert witnesses prepared by the *Royal Statistical Society’s Working Group on Statistics and the Law* (Aitken
et al. 2010). Morrison et al. (2012) provide a comprehensive list of endorsing agencies and institutions, including for example ‘the Board of the European Network of Forensic Science Institutes (ENFSI), representing 58 laboratories in 33 countries’, and the UK and Irish Association of Forensic Science Providers, who have made it part of its standards for the formulation of expert opinion in evaluative forensic science (Association of Forensic Science Providers 2009).

The LR framework is seen as logically correct, since by Bayes’ Theorem a posterior probability – such as ‘it is highly likely the suspect said the incriminating speech’ – cannot be estimated absent a prior probability, to which the expert is not usually privy. In other words, one cannot state the probability that a hypothesis is true, given the evidence adduced in its support, unless one has the probability that the hypothesis is true before the evidence is adduced. This is why the first type of FVC, unless the prior probability is known to the expert, cannot be logically correct. Since a posterior may well impinge on considerations of ultimate issue, which is up to the trier-of-fact to decide, the use of a LR may also be the legally correct option (Robertson and Vignaux 1995).

Apart from its logical correctness, the LR framework has several other important properties. It allows, for example, the combination of evidence of different types, nicely demonstrated in the testing of both automatic and acoustic-phonetic features in hybrid FVC systems. Requiring as it does an estimate of the probability of the evidence under both defence and prosecution hypotheses, the likelihood ratio framework also allows an expert in an adversarial system to be explicitly impartial. Finally – a crucial desideratum in forensic comparison science – the accuracy, and more recently the precision of a LR-based FVC system are also straightforwardly assessed within a LR-based approach (Morrison 2011).

As far as practice is concerned, a LR approach has also been implemented as a matter of course since the mid 1990s in DNA profiling (Forman et al. 2003), although it is not clear that the courts actually understand that the random occurrence probability usually quoted is part of a LR. As reported in Morrison (2009a), the use of LRs in forensic voice comparison was an idea whose time came around the beginning of the new millennium, when its efficacy first began to be demonstrated both with automatic and traditional, usually acoustic-phonetic, features. The results from now well over a decade’s extensive, and continuing, research testing with both automatic and acoustic-phonetic features have shown that the approach works rather well, in the sense that same-speaker speech samples can be rather well discriminated from different-speaker speech samples on the basis of the magnitude of their LRs, e.g. Gonzalez-Rodriguez et al. (2006), Rose (2013). Moreover, Gonzalez-Rodriguez et al. showed in 2007 that the approach can emulate the DNA gold-standard, a
desirable fact, given that ‘DNA profile evidence is now seen as setting a standard for rigorous quantification of evidential weight that forensic scientists using other evidence types should seek to emulate’ (Balding 2005: 55). The LR-based testing of other forensic evidence types is now following: fingerprints (Neumann et al. 2012), handwriting (Hepler et al. 2012) and SMS texting (Ishihara 2012).

Despite these encouraging developments in FVC, which have been interpreted as an incipient paradigm shift towards a maximally objective evaluation of forensic evidence (Morrison 2009a), it is safe to say the idea has not yet caught on. There are probably many reasons for this. The legal profession is, firstly, inherently conservative and thus adheres to the (surely intuitive) assumption that the forensic expert should be asked to use their expertise to estimate the probability of a hypothesis, given the evidence. For example, in a recent draft proposal of standards for the interpretation of forensic evidence, representatives of organizations including the Australian Attorney-General’s Department, the Australian and New Zealand Forensic Science Society, the University of New South Wales’ Expertise, Evidence and Law Program, and the New South Wales Bar Association implicitly upheld the first type of FVC in maintaining that ‘Interpretation [of evidence] includes answering the question as to whether or not … items share a common origin…’ (Standards Australia 2012). Secondly, it appears, many legal practitioners find it difficult to understand the LR approach if they actually encounter it. This can be seen most clearly in inconsistencies in statements concerning likelihood ratio evaluation. For example, the above-mentioned Australian draft standards, in a display of ‘mutually incompatible paradigms of probabilistic reasoning … unable to separate conclusions from evidence …’ (Lucy 2012), proposed a probability-of-hypothesis-given-evidence approach while simultaneously recommending a textbook on the interpretation of evidence (Robertson and Vignaux 1995) which clearly espouses the opposite, probability-of-evidence-under-competing-hypotheses Likelihood-Ratio approach. Balding (2005: 148–154) lists some UK appeal cases where confusion seems to abound. The degree of misunderstanding by the courts is probably best exemplified by the mind-boggling confusion and inconsistencies, described in detail in Morrison (2012a), in the well-known 2010 ruling by the England and Wales Court of Criminal Appeal in R v. T (EWCA 2439, 2010).

It is not just the legal profession that finds the approach difficult to understand, however. A recent book on forensic linguistic evidence misrepresents LRs (Morrison 2009c), and perhaps our best-known phonetics textbook (Ladefoged 2006: 208) defines the LR as ‘the likelihood that the two voices in question are the same as compared with the likelihood that they are different’,
thus confusing it with the prior odds (the same mistake is found in the 2010 *R v. T* ruling). Steam-engine time for the likelihood ratio approach in FVC still looks a very very long way off.

I started to use LR estimates in my case-work in 2002 (some details from actual cases are in Rose 2003, 2006), but the case described here is to my knowledge the first in Australia where LR-quantified speech evidence was actually received in court (there is to date only one other). Australia’s legal system is adversarial, and I appeared for the prosecution. Under the assumption that a ‘logically incorrect conclusion that’s “understood” is no alternative to a logically correct conclusion which needs explanation’ (Berger 2010), this article is intended to document the first FVC case where LR-based speech evidence was received in court in Australia, and I hope thereby to give an idea of what was involved. I will also briefly address areas where the LR estimation might have been improved, given what we know from research in the interim. I also want to use the case to mention what I see as two persisting major problems with LRs, namely conveying the meaning of LRs to a trier-of-fact, and integrating any LR-based evidence with the other evidence in a case.

**Speech evidence and reference sample**

The suspect and offender speech samples in a forensic voice comparison will differ – they always will, because there are always differences between speech samples, even if they have come from the same speaker. *Difference* here is construed in the widest sense. Differences in FVC usually relate to speech sounds, that is, they are phonological or phonetic. But they do not have to be; any difference can be considered, whether linguistic, for example, or non-linguistic. The ensemble of differences noted between the suspect and offender samples constitutes the speech evidence $E_{sp}$. The aim of a likelihood ratio-based approach to forensic voice comparison is to estimate how much more likely one is to get this speech evidence – the differences between the known suspect and unknown offender speech samples – assuming the incriminating speech has come from the suspect (the prosecution hypothesis $H_p$) rather than someone else in the relevant population (the defence hypothesis $H_d$). This ratio of probabilities of evidence under competing hypotheses $p(E_{sp} | H_p) / p(E_{sp} | H_d)$ quantifies the strength of the evidence and is the likelihood ratio.

Estimating a LR is least complicated when the suspect and offender speech samples contain the same material, where *sameness* is defined linguistically. This may be, for example, the same phonemes in the same prosodic environment; or the same word, or words, or even the same phrase. Therefore, when using traditional features to estimate a LR, the first thing to do is to listen to and
transcribe the suspect and offender samples, scouring them for the same material. Although the fraudulent call lasted just over 3.5 minutes, it contained only about 14 seconds of offender speech, and much of that lacked material useful for acoustic-phonetic voice comparison. The offender sample did, however, contain five repetitions of the word yes, and these could be compared with many tokens of the word yes in several recordings of the suspect during previous police and customs interviews and house searches. (It is actually unusual to encounter yes in forensic speech material: people usually say yeah.) In addition, the fraudster’s utterance not too bad was to prove useful in the light of its occurrence, with the same intonation, in recordings of telephone intercepts of the suspect talking to one of his mates. The bulk of the LR estimate in this case thus rested on features extracted from yes and not too bad. I use the term feature for any property of the suspect and offender speech samples that allows the difference between them to be quantified. It can be considered the basic unit of comparison. Features are usually multivariate continuous in nature. In this case, for example, the formant pattern in yes constitutes a feature, represented by measurements of six continuous variables.

**Selection of features**

A very common question at this juncture concerns the choice of features. My reviewers asked: ‘Why didn’t you use other features?’ – specifically, they wanted to know why I chose the features I did over VOT – and why I chose exclusively acoustic features. The same answer can be given to both these questions. Although it is possible to compile a list of the features most commonly compared in FVC (cf. Gold and French 2011), there is actually an infinity of potentially quantifiable differences between speech samples. Therefore, it is impossible to be exhaustive and one needs to choose. One also needs to choose because the infinity of choices is not matched by an infinity of time within which to complete a case. It is another advantage for the likelihood-ratio framework in that it provides a rational way to make these choices. Of the infinity of possible features for comparison, one logically needs those which are liable to give the best strength of evidence – the greatest magnitude LRs. Generally, these are the features which show the first two of Nolan’s (1983: 11, 12) desiderata for speaker comparison features, namely high ratio of between-to within-speaker variance. They are the features that, when combined, are likely to drive the overall LR a long way higher or lower than unity. One of the purposes of the LR-based testing carried out over the last decade is to find which features are capable of giving useful strength of evidence: vowel formants are one. At the time, I was not aware of the strength of evidence to be expected
from VOT, but given the rather large LR I had already obtained from the yes and not too bad features in the case (that has spoiled the story now, but never mind) it would have to have been exceptionally strong to have warranted inclusion – a big LR times a small LR is, after all, still a big LR. (This is not to say that LRs involving VOT are of no use. Morrison et al. (2013) show how it was of use in a disputed utterance case.)

Now to the question of auditory vs acoustic features: dichotomies are only more or less useful (Gould 1987: 8, 9) and I no longer think this is a particularly useful one. In a sense, of course, nearly all acoustic-phonetic features are first and foremost auditory – if we could not identify a vowel by its allophone, we would not be able to measure its acoustics and compare them with those of a comparable allophone. This case contains a paradigm example: it was the intonational pitch of not too bad (an auditory feature par excellence!) which indicated the comparability of F0. More specifically, however, the same argument applies as with the acoustic features: if there are auditory differences that are likely to yield useful LRs then of course they must be evaluated (see Schwarz et al. (2010) for a demonstration of how to get LRs from auditory phonological information). However, there is no point in noting auditory similarities between suspect and offender speech samples that are just as likely to occur if the same speaker is involved as different speakers. Finally, it is not quite correct to say that no specifically auditory features were used in this case: it will be shown below that a LR for the auditory difference between H and L tones on not had also to be calculated.

**Composition of reference sample**

A topic guaranteed to be way more problematic than choice of features is the reference sample. A comparison of the suspect and offender speech samples will allow one to quantify the difference between them, but in order to estimate the probability of getting this difference between the offender sample features and another person chosen at random from the relevant population, a reference sample is of course necessary. The notion of relevant sample (or relevant population or background sample or background population) in forensics is a vexed one. The most useful characterisation is that found in Aitken and Taroni (2004: 281), namely the population defined by a combination of defence hypothesis and background information. In this case, no explicit alternative hypothesis having been nominated by defence, it was sensible to assume from the way it sounded that the offender voice belonged to another adult male speaker of General Australian English, and I collected a sample of 35 adult males with ages uniformly distributed between early 20s and early 70s who
agreed to be phoned up and recorded. One of my referees was worried about
the effect of this age range on formants. While it would be possible to test the
reference sample for this, it would be beside the point: in the absence of
information to the contrary, it would certainly have been wrong to narrow the
reference sample down. (It is also not necessarily a good defence strategy: a
more narrowly defined reference sample may well increase the LR denomi-
nator, and thus decrease the LR, but it will also change the prior odds in an
unwanted direction.)

Unlike, for example, *I’ll blow yer fuckin’ head off*, the test material in this
case (the suspect and offender data) has the enormous benefit of easy ecologi-
cally valid elicitation: reference sample speakers – my mates and colleagues –
were instructed to appropriately reply *not too bad* and *yes* or *no* to a set of
questions I asked, such as *how’s it going?*, *what was your week-end like?* or
*are you inside?*, *is it raining?* I tried to indirectly prime them once at the
beginning of the elicitation simply by saying *not too bad* with the correct
intonation, but they were not explicitly told the correct intonation to produce,
and neither were they corrected if they said *not too bad*, as some sometimes
did, with an undesired intonation. (Some tokens, for example, were said with a
falling nuclear tone on *bad* (L.M.HL), or with a low pre-head on *not* and a high
head on *too* (L.H.LH). It was also common for speakers to prefix a prolonged
low back vowel, probably representing ‘oh’, as in *oh not too bad.*) In this way
it was possible to obtain several replicates of *yes*, and *not too bad* with the
correct, natural sounding intonation, from most speakers in the reference
sample. To sample non-contemporaneous within-speaker variation, subjects
were phoned and recorded on several occasions over the course of several
weeks.

One of my reviewers worried that, since the situation in which the speakers
were recorded was cognitively different from that of either the suspect or
offender recording, this could have had an unpredictable effect on the acous-
tics. Of course one is never, ever, going to be able to mirror the precise situation
of the actual test data with the reference sample. One reason is that, as Beckman
(1997) notes, we do not understand all of what is going on in such exchanges:

The prosody of spontaneous speech is affected profoundly by the social
and rhetorical context of the recording, and these contextual factors can
themselves vary widely in ways beyond our current understanding and
control …

However, other than me suborning a 35-fold telephone fraud, the social con-
text was controlled in so far as it was an informal conversational interaction
between two mates, and the elicited utterances certainly sounded convincing
enough to dispel worries about play-acting. (An unexpected twist to the case was that the fraudulent telephone call turned out to be just that – play-acting: the JP Morgan employee taking the call was also part of the plot.) The rhetorical context, however, was not controlled as well as it might have been. This is because some of the not too bad were said in response to questions other than an opening how’s it going? I suspect that individuals actually show less variation from occasion to occasion in their intonational F0 for such phrases than I obtained, and therefore the within-speaker variation was probably overestimated. This would have led to greater magnitude LRs than I obtained, which would in turn have disadvantaged the defence. It is good that it did not happen, therefore.

In the sections below, I will exemplify some of the features used in the forensic voice comparison and describe how their LRs were estimated.

**Formant pattern in yes**

Like any artefact, the sound radiated from a speaker’s head carries the imprint of its origin. In speaking, the origin is twofold: the vocal tract in which the sound was generated, and the instructions to the vocal tract for realising the speaker’s communicative intent. To the extent that speakers differ in the constraints imposed by their vocal tract size and condition, and the choices they make in implementing their communicative intent with it, there will be between-speaker differences in the features we can extract from the speech acoustics of the voice. But there are limits to the amount of between-speaker variation in tract size and use and thus there are limits to between-speaker variation in radiated acoustics, and it is relatively easy to find two speakers with very similar acoustic values for a given speech sound, or even a set of sounds. In addition, there is also variation within speakers: it is a phonetic truism that no-one ever says the same thing in exactly the same way. Generally, however, between-speaker variation is greater than within-speaker variation, and this, together with the fact that we do not all differ from each other in the same way – my *ee* vowel might have different acoustics from yours and similar acoustics to Bruce’s, whereas my *ah* vowel might have similar acoustics to yours and different acoustics from Bruce’s – makes forensic voice comparison possible. There are, therefore, differences between the suspect and offender speech samples. The idea is to be able to say whether these differences are more likely to be of the type one finds between speakers, or within.

Speech samples are commonly compared forensically with respect to the centre frequencies of their vowel formants. The reason is that formants, being vocal tract resonances, are perhaps the clearest instance of relatively easily
quantifiable speech acoustics reflecting the size and shape of the vocal tract that produced them. As far as formant centre-frequencies are concerned, the relationship between vocal tract and acoustic output is shown in the well-known formula at (1) for the resonant frequencies $F_n$ in Hz for a vocal tract of uniform cross-sectional area and length $l$ in cm closed at one end. This formula says that the frequency of any given formant with such a vocal tract is determined by two things: the speed of sound ($C$) and the length of the vocal tract ($l$). $C$ can be conventionally taken as a constant at 35,000 cm/sec. Since $l$ is equivalent to the length of the supralaryngeal vocal tract, a schwa coming from a 17.5 cm long vocal tract, for example, will have a second formant frequency $F_2$ of 1500 Hz.

$$F_n = \frac{(2n-1)C}{4l}$$
where $C =$ speed of sound in cm/sec.

Vocalic F-pattern has been the most tested feature within a likelihood ratio-based approach to forensic voice comparison. This likelihood ratio-based testing has investigated both the validity of forensic voice comparison with vowel formants, and what strengths of evidence can be expected from different kinds of vowels, and has demonstrated that LRs can indeed discriminate between speech samples from the same speaker and speech samples from different speakers (the crucial question in FVC); but that the strength of evidence varies with different vowels. Diphthongal F-pattern has been found generally to yield the best strength of evidence (Rose 2010b), probably because there is information from two targets, and, as anticipated by work trialling speaker characterisation with formant trajectories (McDougall 2006, McDougall and Nolan 2007), and demonstrated by Morrison (2009b), there is a greater potential for between-speaker differences in getting from the first target to the second. The diphthongal F-pattern of yes was shown in Simmons (1997) to be promising for LRs. Thus the presence of 6 tokens of yes in the offender speech sample and 23 tokens in the suspect, all said as an affirmative response to a question, meant that a comparison of suspect and offender F-pattern in the /je/ of yes was likely to yield a useful LR (i.e. one deviating considerably from unity and therefore useful for distinguishing between prosecution and defence hypotheses).

Both suspect and offender yes tokens sounded to have a more centralized /e/ than normal, and lower than normal pitch on /s/ (reminiscent of the pharyngealised /s/ in Arabic). Figure 1 shows spectrograms of yes tokens from offender and suspect. The offender token shows a clear zero at 4 kHz and typical bandlimiting effects from the phone, although the presence of clear
energy just below 4 kHz shows that its upper bandpass is unlikely to affect the F3. Both show unremarkable F-pattern configurations given the auditory impression – note the lower than normal cut-off frequencies for the /s/.

![Figure 1: Spectrograms of yes tokens from offender (left) and suspect (right). X axis = duration (csec.), y axis = frequency (Hz).](image)

The offender and suspect /je/ F-pattern was sampled at onset, mid-point and offset. Figure 2 shows the time course of the /je/ F-pattern in the suspect’s 23 and offender’s 6 replicates, to the extent that it can be represented by three sampling points. It can be seen that that the sampled F-pattern values of the suspect and offender are fairly similar, the offender’s values lying almost totally within the distribution of the suspect.

![Figure 2: Equalised time course of /je/ F-pattern in yes for suspect and offender replicates. Dark lines = offender. X axis = sampling point (1 = onset, 2 = mid-point, 3 = offset).](image)
Figure 3: Probability densities of offender F-pattern in yes /je/ (light shade) compared with suspect (dark shade, outlined). X axis = frequency (Hz), y axis = probability density.

Although phoneticians are probably most used to the kind of plot shown in Figure 2, a way of demonstrating similarity that is more relevant to LR estimation is with probability densities, and this is given in Figure 3, which plots the densities of F1, F2 and F3 at the three sampling points of /je/ for suspect and offender. Again, Figure 3 shows that the offender’s samples are similar to those of the suspect in that they generally show a certain amount of overlap with the suspect’s.

**Quantifying similarity**

One term of the likelihood ratio relates to the probability of getting the offender values, assuming they have come from the suspect. This is sometimes called the *similarity term* because it quantifies the degree of similarity between suspect
and offender samples. The similarity between suspect and offender obtains between their sample means, as one needs to know the probability of getting the difference between the mean of the suspect data and the mean of the offender data, assuming they have come from the suspect. Therefore, it is necessary to invoke the error of the sampling distribution of the difference between means. If one repeatedly takes two samples, with replacement, from the same population with a variance $\sigma^2$, one with sample size $= n$ and one with a sample size $= m$, and finds the difference between the two samples’ means, the distribution of these differences will be normal around a mean of 0 and with standard deviation of $\sqrt{\sigma^2/n + \sigma^2/m}$. Of course we do not know the value of the suspect’s population variance $\sigma^2$, and this has to be estimated. One way of doing this is to use the suspect’s sample values; another is to use an estimate from the reference sample (this is the method actually used in the approach described below); yet another method is to estimate the suspect variance from a combination of suspect and reference sample values, weighted to reflect the amount of suspect data (Ramos-Castro et al. 2006).

Figure 4 illustrates the probability of getting the difference between the suspect and offender data means, assuming the offender’s data have come from the suspect (the suspect’s variance has been estimated from their sample, under the assumption that its values are normally distributed). In Figure 4 is shown, for each sampling point of each formant, the distribution of the difference between means assuming one sample of size 6 (for the offender) and one of size 23 (for the suspect), and a standard deviation corresponding to that of the suspect sample. The top left panel, for example, shows that the difference between the means of the suspect and offender F1 onset samples – it was c. 17 Hz – would have a high probability density (c. 0.01) if the offender sample had come from the suspect. The top right panel, on the other hand, shows that the probability density for the difference between the suspect and offender mean F1 offset values – it was 52 Hz – would be very low assuming provenance from the suspect. It can be seen from Figure 4 that seven of the nine probability density values are high or average, which means that the probability of getting the offender yes F-pattern data, assuming that they have come from the suspect, will be reasonably high.

Quantifying typicality

In estimating the strength of evidence with a likelihood ratio, however, similarity is only half the story. One also needs to know the probability of getting the offender value assuming it has come from someone other than the suspect
Figure 4: Notional evaluation of similarity between suspect and offender *yes* /je/ F-pattern: difference between suspect and offender *yes* mean F1 to F3 at all three sampling points (vertical lines) plotted against sampling distribution of difference between means given suspect provenance. X axis = difference between suspect and offender means (Hz), y axis = probability density.

chosen at random from the relevant population – in other words, one also needs to know how typical the offender value is. This is where the reference sample comes in. For a likelihood ratio to be useful to the prosecution, the suspect and offender samples must be more similar than they are typical. *Mutatis mutandis* for the defence. Figure 5 shows the /je/ F-pattern for the 31 individual reference sample speakers used. There are some obvious between-speaker differences in F-pattern range, the extremes being speaker 27 with the highest overall F-pattern, and speaker 25 with the lowest. Some between-speaker differences in trajectory (as far as can be seen from three sampling points) can also be discerned.
Figure 5: Individual equalized-duration F-patterns of /je/ in yes for reference sample speakers. X axis = sampling points (1 = onset, 2 = midpoint 3 = offset). Y axis = frequency (Hz) (cont.).
Figure 6 plots the duration-equalised time course for the reference sample /je/ F-pattern means together with the mean values for the suspect and offender. It can be seen that the between-speaker variance in the reference sample F2 and F3 is quite big, but more importantly it shows that the F-pattern in the suspect and offender, apart from their onset F2 values, is not very typical. In particular, both suspect and offender samples have unusually low offset values. This presumably correlates with the audibly lower, more centralised nucleus heard in both.
The combination of relative high similarity between suspect and offender values compared to their low typicality compared to the reference sample means that a LR greater than unity is to be expected: one would be more likely to observe the difference between suspect and offender values assuming they had come from the same speaker.

**Yes F-pattern likelihood ratio**

How much more? In order to find this out, a LR was estimated using the Aitken and Lucy (2004) multivariate likelihood ratio (MVLR) formula. This rather intimidating formula is given, and some of its terms explained, in Rose (2013: 94–95). Its performance has been tested in many studies. It calculates a LR based on the difference between the suspect and offender means when evaluated against estimates of within- and between-speaker variance derived from the reference sample, also taking into account the number of suspect and offender replicates (because the more replicates the stronger the evidence, *ceteris paribus*). Its most important feature is that it is able to take correlation between variables into account, by exploiting their multivariate nature. (In this case the potential multivariate data are the F1 F2 and F3 values sampled at the three sampling points.) This is very important. In a /je/ F-pattern, for example, one would expect correlation between formant values at the different sampling points, and also perhaps correlation between all three formants, if they have come from the vocal tract of the same speaker. Ignoring correlation between variables will usually lead to overestimated LRs (Rose 2010a).

There are two versions of the MVLR formula, depending on whether one assumes normality for the reference sample. If not, the reference sample distribution is estimated with a kernel density. A LR for the F2 and F3 values at all
three sampling points in /je/ was obtained. Only F2 and F3 values were evaluated as it was assumed that the difference between suspect and offender F1 might have been compromised by the telephone transmission. Both normal and kernel density LRs (72.1, 76.0 respectively) indicated that the difference between suspect and offender /je/ F-pattern, thus quantified, was about 70 times more likely had they come from the same rather than different speakers.

**Fundamental frequency in *not too bad***

The second feature used in the forensic voice comparison was the fundamental frequency in the expression *not too bad*. Fundamental frequency (F0) is the main acoustic correlate of pitch and is the acoustic reflex of the rate of vibration of a speaker’s vocal cords: the higher the vibration rate, the higher the perceived pitch. Just as with formant centre-frequencies, F0 can also be related to the vocal tract that produced it, and formulae relating acoustics to vocal tract in the case of F0 are given at (2) and (3). The rate of vibration of a speaker’s vocal cords is partly determined by their length. As shown in the formula at (2) F0 is inversely proportional to cord length, so that a speaker with relatively long vocal cords will have lower F0 values, other things being equal. F0 is also determined by cord mass. The formula at (3) shows F0 is proportional to the square root of the ratio of stiffness to mass, so other things being equal a greater mass will correspond to a lower F0.

\[
F0 = \frac{1}{2L_m} \sqrt{\sigma_c / \rho}
\]

where \( L_m \) = the length of the membraneous cords, \( \sigma_c \) = the longitudinal stress in the cord cover (tension in the cords divided by the cross-sectional area of vibrating tissue), and \( \rho \) = the tissue density (a constant, at about 1040 kg/m\(^3\)) (Titze 1994: 200)

\[
F0 = \frac{1}{2\pi} \sqrt{k/m}
\]

where \( m \) = vocal cord mass, \( k \) = stiffness (Titze 1994: 193)

To the extent that speakers differ with respect to cord mass and length, their F0 will also differ, other things being equal. The simplest example is the sex-related difference in F0 between males and females. Males generally have longer, more massive cords, and hence lower F0. Of course, this is not, by any means, the whole story, and, like anything else in speech, a speaker’s F0 is a long way from invariance. Exactly the same limiting considerations apply as
with formants. There are limits to the extent of between-speaker variation in vocal cord length and mass, and if two different speakers have similar sized cords, their F0 will also be expected to be similar. Equally important is that, although the rate of vibration of the vocal cords is determined anatomically, there is a very large number of factors that influence this vocal cord physiology in the first place. F0 plays an important part in signalling linguistic, paralinguistic and non-linguistic information, reflecting the choices speakers deliberately make in their communicative intent. For example, pitch is used to signal a wide range of linguistic information within the broad suprasegmental categories of tone, stress and intonation, and the F0 of a speaker of a language such as English will differ depending on whether they are saying *not too bad* with a high pitched tone on *not* or a low, or whether they are saying *bad* with a rising pitched tone or a falling. Variation in F0 is constrained by a bewildering number of factors, including, for example, psychological and physical health, intoxication, age and smoking habit – more can be found in Braun (1995).

The usefulness of F0 as a forensic feature is limited if these factors are not controlled, as otherwise the amount of within-speaker variation in F0 will approach that of between-speaker variation. A small amount of research has shown that, if these factors can be adequately controlled, F0 can be of use in discriminating between samples from the same speaker and samples from different speakers with likelihood ratios. Kinoshita et al. (2009), for example, showed that long-term F0 distributions could be used to discriminate rather well between same-speaker and different-speaker speech samples with LRs, but the amount of speech necessary was truly long term (between 10 and 25 minutes) and forensically unrealistic. Zheng and Rose (2012) demonstrated the same discriminatory potential of LRs using F0 distributions from much shorter amounts of running speech in Cantonese (between about three-quarters and one-and-a-half minutes).

In this case, there was insufficient offender speech data for even a short-term F0 comparison, but near identity in intonational structure – perhaps not a common occurrence – made a F0 comparison possible. This was because all test data *not too bad* s are said in response to the interlocutor’s asking how the suspect/offender is. The *not too bad* s have very similar intonational structure typical for this conversational interchange, conveying the typical rise nuclear tone meanings of supportive interest encouraging further conversation (Wells 2006: 41, 91). They have a rising nuclear tone on *bad*, realised with a low rising allotone. The offender *not* carries a high head realised with a high, slightly falling pitch (probably induced from the coda stop); the suspect’s *not* has either a high head realised in the same way, or a low head, with a low pitch. The pitch on *too* is interpolated in the expected way. This near identical intonational
structure in suspect and offender *not too bad* means that the F0 values realizing the structure are highly comparable and might be expected to yield a useful likelihood ratio (i.e. one that deviates substantially from unity).

Figure 7 shows the F0 realising the [H.L.LH] intonational pitch of the offender, aligned with its wideband spectrogram. F0 on *not* can be seen to drop from about 200 Hz to 175 Hz, whence it drops further on the nucleus of *too* to about 125 Hz. The F0 shows a small c. 15 Hz increase from its minimum value of 125 Hz in the /b/ hold, and rises on the nucleus of *bad* with a slightly convex contour from about 145 Hz to peak at about 185 Hz.

![Figure 7: F0 time course of offender's *not too bad* (right-hand scale in Hz) superimposed on wide-band spectrogram (left-hand scale in Hz). X axis = duration (csec.)](image)

The degree of similarity between suspect and offender *not too bad* F0 can be easily conveyed graphically by plotting the offender values against the distribution of the suspect, as shown in Figure 8. In Figure 8 the offender’s *not too bad* F0 has been plotted with a thick line and the *not too bad* F0 from the suspect’s 15 replicates is shown with thin lines. The duration has been equalised using the onset of /ɛː/ in *too* and the peak F0 in *bad*. It can be seen that, as far as the F0 on *too bad* is concerned, the similarity is considerable, with the offender’s F0 time course lying completely within, and in some places almost exactly in the middle of, the suspect’s distribution.

The F0 on *not* is different. The offender F0 on *not* appears to lie almost at the top of the suspect distribution, similar to a few but quite a long way from most of the suspect’s *not* F0 values. The reason for this is that the suspect said *not too bad* with both H and L tones on *not*, whereas the offender’s single *not too bad* had H. The tone on *not* is indicated for the individual replicates in the figure. This is an example, probably not rare in speech, where there is both a categorical difference (between H and L on *not*) and a continuous difference (between F0 means on *not*) to evaluate.
Figure 8: Suspect (thin lines) and offender (thick line) *not too bad* F0 plotted as function of equalized duration. X axis = duration equalized around /ɛ/ onset (0%) and peak F0 in *bad* (100%). Y axis = F0 (Hz), H, L = high, low tones on *not*.

Figure 9: Examples of within- and between-speaker variation in reference sample: three replicates of *not too bad* F0 from two speakers on two different occasions. X axis = duration (sec. from peak F0 in *bad*), y axis = F0 (Hz). H/L = F0 on *not* with high/low tone.

The *not too bad* F0 values from the 35 reference sample speakers used showed no surprises in variance structure. Figure 9, which plots the F0 time course for just two of the reference speakers on two different occasions, illustrates this. There was, firstly, expected within-speaker variation in phonological
structure between a H and a L tone on *not*, just as in the suspect. The first speaker in Figure 9 displays this behaviour: in recordings 1 and 2 the different F0 height on his *not* corresponding to the H vs L distinction can be easily seen. In his first recording, two of his tokens had a H tone and one L; the pattern reversed in his second recording. This speaker also shows considerable non-phonological variation in F0 between sessions: his *not too bad* in recording 2 are much higher (they sounded more enthusiastic than in recording 1). A comparison with the second speaker in Figure 9 illustrates typical aspects of between-speaker variation. Most obviously, he shows some different overall F0 values from the first speaker. But the two speakers also differ in their within-speaker variation: unlike the first speaker, the second speaker always had a H tone on *not* and shows little difference from recording to recording.

I decided to parametrise the F0 time course in *too bad* by sampling at four points: mid point of the vowel in *too*, and at first target, peak, and mid-way between targets in the vowel in *bad*. Because of the aforementioned associated phonological variation between H and L, F0 on *not* was evaluated separately. The resulting reference sample is shown in Figure 10.

![Figure 10: Individual F0 time course of *not too bad* for reference sample speakers. X axis = sampling points (1, 1.5, 2 = onset, mid-point, offset F0 in */o*; 3 = */æː* mid-point F0; 4, 5, 6 = onset, mid-point, peak F0 in */æ*/. Y axis = fundamental frequency (Hz) (cont.).](image)
Figure 10 (cont.)
Figure 11: Typicality of suspect and offender *too bad* F0 time course: time course of suspect and offender mean values (dark lines) compared against mean values of reference sample speakers (grey lines). X axis = sampling points (1 = /æː/ mid-point F0; 2, 3, 4 = onset, mid-point, peak F0 in /æː/).

Figure 11 shows the typicality of the mean suspect and offender F0 time-course in *too bad*. It can be seen that both suspect and offender values share an atypical trajectory for the realisation of /LH/ on *bad*, which lacks the concavity evident in most of the reference sample. This means that although F0 values in *too* and at peak of *bad* lie within the distribution of the reference sample, onset and mid-point values are atypically high. The distribution of the reference sample peak also appears bimodal, in which case the suspect and offender peak values may lie atypically low.

**Not too bad** F0 likelihood ratio

The four sampled *too bad* F0 values were treated as multivariate data and LRs were estimated with the MVLR, again modelling the reference sample both normally and with a kernel density. The difference between the suspect and offender F0 values in *too bad* was estimated at about 20 times more likely assuming they had come from the same speaker, irrespective of the normal (LR = 21.7) or kernel (LR = 20.6) modelling of the reference sample (so the abovementioned apparent bimodality of the reference peak values did not make any difference). Given the high degree of similarity between suspect and offender samples demonstrated in Figure 8, this relatively low LR value for the comparison is salutary. Given the relative atypicality of the suspect and offender values, the low LR value is probably a function of the generally high within-speaker variation typical for F0, even for a set phrase like *not too bad*.

The difference between the suspect and offender replicates in tone on *not* meant that F0 values on *not* had to be evaluated in a different way. One way to do this is to estimate both a categorical and continuous LR. Categorically, only
half of the suspect’s not tones were H, which means that there is a 50% probability of observing the offender’s H assuming it had come from the suspect. The incidence of H in the reference sample ranges from 0% to 100%, but generally it is quite high, as most reference sample speakers said not with H. The mean incidence in the reference sample is 75%. This means that the best estimate of the probability of observing the offender’s H assuming it had come from someone other than the suspect in the reference sample is 75%. The categorical LR for the H tone on not is then (50/75 =) 0.67, meaning that one would be (1/0.67 =) 1.5 times more likely to observe the offender’s H on not assuming it had come from someone other than the suspect. Now having taken the difference between the H and L tones into account, we can evaluate the continuous LR based only on the H tone F0. That is, a LR is estimated by evaluating the difference between the offender’s H tone not F0 and the mean F0 of the suspect’s seven H tone not replicates against the F0 of the H tone nots in the reference sample. The value for this continuous LR is c. 1.7 – a little more likely under the prosecution hypothesis than the defence. Combining the categorical and continuous LRs for the intonational features of not gives a combined LR of (1.7 * 0.67 =) c. 1.1, manifestly of no use.

Another way of approaching this is simply to ignore the linguistic difference between H and L tone on not, and evaluate the difference between the mean of all the suspect’s not replicates and the offender’s not against all reference sample replicates, irrespective of their tone. If this is done, the LR is c. 1.45 – a little more likely assuming same-speaker provenance, but basically very nearly useless. Whether one takes the linguistic information into account or not, therefore, makes no difference in this case: both ways of estimating the LR for the F0 difference between suspect and offender in not result in equivocal LR values, and can be ignored. The overall LR for not too bad F0 thus remains at about 20.

Other features

All segmental aspects of the not too bad acoustics are also theoretically comparable, and LRs for the F-pattern on its three constituent vowels (/o/ in not, /ʌː/ in too and /æ/ in bad) were also estimated. These are not particularly good vowels as far as expected strength of evidence goes. Rose (2010b: 32) shows that out of the 18 Australian English vowel phonemes they rank eighth (/æ/) tenth (/o/) and seventeenth (/ʌː/). As expected, comparison with /o/ was complicated due to nasal poles associated with quasi-Helmholtz and nasal cavity resonances induced by /n/. Figure 12 shows the similarity and typicality for these three vowels. In the left-hand column the F-pattern time course of the
Figure 12: Similarity and typicality of suspect and offender F-pattern values in /o/ (top row), /æː/ (mid row) and /æ/ in *not too bad*. Left column shows offender’s F-pattern time course (thick line) plotted against suspect’s values (thin lines). Nasal poles in /o/ are not shown. X axis = equalised duration (%*0.1), y axis = frequency (Hz). Right column shows target F-pattern centre-frequency means for offender (vertical dotted lines) and suspect (vertical solid lines) plotted against the reference sample distributions. Fn = nasal pole. Y axis = probability density, x axis = frequency (Hz).
suspect and offender values is compared, and in the right the suspect and offender means for F-pattern target values are compared against the reference sample. It can be seen that for all three vowels there is, as with the F0, considerable similarity, the offender values lying almost totally within the distribution of the suspect’s values. Most of the values at target can also be seen to be fairly typical.

The F-pattern differences were parametrised by mean values taken over quasi-steady-state portions at vocalic target, and point measurements at offset. As in yes, some values were discarded because of possible influence from the telephone transmission. The resulting MVLR estimates were all greater than unity: c. 24, 5 and 11 for /oːæ/ respectively.

It was noted above that both suspect and offender /s/ in yes showed reasonably high amplitude narrowband noise cutting-off at about 3.0 kHz. In addition to the not too bad vocalic F-pattern, I attempted a crude LR estimate for this low spectral cut-off by inspecting all same-speaker and different-speaker pairs in the reference sample, and estimating how many agreed in a low cut-off in /s/ (no measurement was involved, just eyeballing). This showed a low cut-off to be conservatively about 2.5 times more likely in both samples if they had come from the same rather than different speakers.

**Overall result and caveats**

A naïve Bayes combination of the LRs from all the features described (F0 and F-pattern in not too bad, F-pattern and low cut-off in yes) yielded an enormous overall LR of about 11 million (one feature not described contributed an additional LR of c. 2). Enormous magnitudes are what often happens with naïve Bayes, and this value will almost certainly have over-estimated the strength of evidence, since a naïve combination (taking the product of the individual LRs) assumes that the features being combined are not correlated. Correlation between variables within features (i.e. within-segment correlation) is pretty much certain: the c. 160 speakers of the (1967) Bernard data set show a c. 50% correlation between F2 and F3 in /æ/, for example, and a c. 45% correlation between F2 and F3 in /ɛː/; and F2 and F3 in /j/ will also be strongly correlated. These within-segment correlations will be handled by the Mahalanobis distance of the MVLR formula. Correlation between features, however, will not be. There was clearly going to be some between-feature correlation in not too bad. One would expect tongue height assimilation in the unstressed /ɛː/, for example, to occur between the two low vowels /o/ and /æ/. But even discounting these stress effects, experimentation of the type demonstrated on separate vowels in Rose (2010a: 322–323) showed correlations.
significant both in frequentist and Bayesian terms. There is a c. 40% correlation between /æː/ F3 and /æː/ F2, for example. Since I suspected some between-feature correlation, then, but could not estimate its effect on the LR, I simply discarded the putatively correlated LRs (e.g. both formants in /æː/) to arrive at a much smaller LR estimate of c. 300,000. According to Bayes’ Theorem, a LR of this magnitude means that there would now have to be more than roughly 16,000 others, including the suspect, who could have said the incriminating speech before the posterior probability that the suspect said it fell below 95%.

It is sensible not to take even this lower estimate at face value, however. This is not a simple problem in aleatory probability: we do not know, and probably never will ‘the physical nature of the system generating the uncertainty in outcome’ (Lucy 2005: 17). Although it is able to take many things into consideration, the two-level MVLR formula still has to make a lot of simplifying assumptions, especially with respect to variance estimation. The acoustic-phonetic commonplace, that speakers can and do differ in their variance, was illustrated in Figure 9. There are two ways in which the MVLR simplifies this complex variance structure. Since it is a two-level random effects model, its estimate of the within-speaker variance, taken as it is from a speaker’s tokens pooled over both sessions, includes but does not distinguish the separate component of between-session variance. Neither does it take into account between-speaker differences in variance. Rather, it estimates the between-speaker variance from the mean variance across all speakers in the reference sample. Thus the sampling distribution of the difference between means illustrated in Figure 4 would be based on the mean within-speaker variance of the reference sample, not the variance of the suspect. There are good reasons for these simplifications, of course. One is that, because offender and suspect samples may be sparse, the approach is more robust. Another consideration is that the sample size (of both reference speakers and suspect and offender replicates) would not have been big enough to justify confidence in an exact value. This is in turn because the sample size needed to properly estimate a multivariate density – that is, with an acceptable error – increases geometrically with the number of dimensions (Silverman 1986: 93–94, Rose 2013: 98).

Because of these considerations, I think it is best, if possible, to try to avoid precise figures, substituting them with a verbal LR statement such as ‘very very much more likely to get the differences between suspect and offender recordings assuming they have come from the same speaker’. The weakness of this of course is its circularity. Anyone can sensibly retort: ‘And just what do you mean by very very much more?’ To which the only proper answer is the number. At the very least, one can say that the LRs for all of the features examined pointed in the same direction.
Critical hindsight

The now five years since this case went to court have seen continuing improvements in traditional FVC LR estimation. The most important ones have been of three main types and have involved recruiting methods from automatic forensic speaker recognition. The first type has been improvements at the most basic level, in features and their parametrisation. It has been shown that improved performance – that is stronger LRs and better discrimination – is achieved by parametrising formants with features that model their trajectory, like discrete cosine transforms or polynomial coefficients, rather than with measurements at certain points such as presumed vowel targets (Morrison 2009b). It has also been shown that using spectral shape parameters like cepstral coefficients vastly improves performance when vocalic segments are compared (Rose 2013), and that spectral shape parameters can also be of use with segments that would otherwise be difficult to quantify, like fricatives (Rose 2011).

A parallel improvement has come at the metalevel of system evaluation, with ways of quantifying accuracy and precision5 (Morrison 2011). Most important have been the use of logistic regression (Pigeon et al. 2000) to calibrate distance measures within a feature, or scores, and to combine evidence from separate features, the latter providing a potential solution to the problem of possible between-segment correlation (Morrison 2012b). Explained examples of many of these improvements (within-feature calibration, between-feature fusion, spectral shape parameters and evaluation of validity) may be found in Rose (2013: 96–99, 103–105).

All of these improvements would probably have helped enormously in this case. Firstly, given the segmental and suprasegmental identities in not too bad and yes, it would have obviously been advantageous to parametrise both F-pattern and F0 trajectories with coefficients, rather than use point estimates. This would definitely have made the quantification easier and quicker and, because it would have enabled me to compare the F-pattern and F0 over all of their time course, rather than at a few sampling points, probably would have resulted in LRs of greater magnitude. Secondly, the whole of the /s/ spectrum in yes could have been compared using cepstrally-mean-subtracted cepstral coefficients, rather than just the fact that it had a low cut-off.

Most importantly, the problems with handling possible between-feature correlation might have been avoided by logistic-regression fusion of the LRs from the different segments, and features that I had to jettison on suspicion of correlation could have been retained, thus allowing all the evidence quantified to have contributed to an outcome. It is worth illustrating this final point here,
because it also relates to an important point about demonstrating the validity of the system.

**Demonstrating system validity**

Imagine that we have suspect and offender speech samples from which we can extract the two features of *yes* F-pattern and F0 in *not too bad*. The suspect and offender samples both contain five tokens of *yes*; the suspect sample contains five tokens of *not too bad*, the offender four. Figure 13 compares these values: it can be seen that their mean values (shown with thicker lines) are fairly similar.

![Figure 13: Illustrative suspect and offender samples compared for F0 in *not too bad* (left) and F-pattern in *yes*.](image)

The suspect and offender values are then set aside and a FVC system built to determine to what extent, if at all, it is possible to distinguish between same-speaker and different-speaker samples using the selected F0 and F-pattern features. This is done empirically by recording a group of speakers on different occasions. It is important that the number of replicates in the system parallel the suspect and offender replicates, thus both sets of *yes* recordings must contain five replicates per speaker, and one set of *not too bad* recordings must contain four and the other five. Each speaker’s recording on one occasion is then compared with themself on the other, and with the recordings of all other speakers. For each comparison a score is obtained with the multivariate LR formula. This gives a set of scores from same-speaker comparisons of each feature, and a (much larger) set of scores from different-speaker comparisons for each feature. The Tippett plot for the same-speaker and different-speaker scores for the *yes* F-pattern and *not too bad* F0 is shown in the top panel of Figure 14. As they were obtained with a relatively small number of speakers,
cross-validation by a leave-one-out method was used. As can be seen, both features have fairly similar scores, and, as is usual for features with a large number of variables, are shifted to the left, with some badly calibrated same-speaker comparisons (i.e. bad counterfactual values). These same-speaker and different-speaker scores are then analysed with linear multivariate logistic regression the coefficients of which allow us to convert a score into a likelihood ratio. The logistic-regressive coefficients (also termed weights) are shown in the top panel of Figure 1 and comprise slope coefficients for the two features, and an intercept. The slope magnitudes show that both features contribute fairly equally to the outcome. The scores from the two features are now fused to obtain a single pair of LRs for the combined two-feature system. The resulting LRs are shown in the bottom panel of Figure 1, where the typical shifting and scaling effect of the fusion can be seen, resulting in the reining-in of the bad counterfactual scores. This plot shows the strength of evidence that can be expected from our two-feature system.

The Tippett plot at the bottom of Figure 1 is normally taken to be a graphical demonstration of the validity of the system, to be shown to any interested party, especially of course a trier-of-fact. As explained in Rose (2013), there are two ways of making sense of such plots, one discriminatory and one information-theoretical. The Tippett plot in Figure 1 suggests that a system using *yes* F-pattern and *not too bad* F0 values in the way described can discriminate between non-contemporaneous same-speaker and different-speaker speech samples with an error rate of about 5% for different-speaker comparisons and about 3% for same- (out of the 528 different-speaker comparisons it evaluated 28 counterfactually (i.e. with a logLR greater than 0), and out of the 33 same-speaker comparisons one was wrong). The sensitivity-specificity test, a simple and common metric from medical diagnostics, summarises this result. This is the probability of getting a correct same-speaker decision relative to the probability of getting an incorrect different-speaker decision. This statistic, which is also a likelihood ratio, is discussed and exemplified in Rose (2013: 99). Its value for this system is thus (97%/5% = ) c. 19, which means one is about 19 times more likely to get a log10LR greater than 0 if the samples came from the same speaker than from different speakers.

This discriminatory construal is relatively easy to understand, but does not strictly speaking represent the true nature of LRs *qua* strength of evidence indicators. The crucial datum here is the log likelihood-ratio cost (Cllr). This is an information-theoretical construct which is currently used to evaluate the performance of LR-based detection systems (Brümmer and du Preez 2006). The Cllr has to be below unity if the system is providing any information, and ideally it should be well below unity. The Cllr of 0.15 is respectably low, but
it might be difficult to explain why this is so to the poor trier-of-fact (the formula is a little complicated, and the theory behind it considerably so), and there is no scale which states a cut-off value above which a system is still acceptable. In addition, the CIlr is relative and can only be used to compare systems operating on the same data. It would therefore be of limited use to a trier-of-fact.
If it is agreed, somehow, that the system is sufficiently valid, the suspect and offender data illustrated in Figure 1 can then be evaluated in its terms and a LR estimated for the evidence. Separate scores are obtained for the comparison between the suspect and offender yes F-pattern and not too bad F0 by evaluating the difference between suspect and offender means against a reference sample with the MVLR (which is now understood to simply output scores). These scores are 6.31 (F0) and 0.131 (F-pattern). From them, and the logistic-regression weights, the LR for the two features combined can now be estimated as at (4), after the appropriate conversions to and from natural logarithms. The resulting LR of 10.6 indicates that one would be about 10 times more likely to get the combined difference between suspect and offender features if they had come from the same speaker rather than different speakers. Not a great strength of evidence to be sure – with a LR of 10.6, even if there were only one other in addition to the suspect who could have said the incriminating speech, the posterior probability that the suspect did in fact say it can only reach about 91%. But at least the LR is not counterfactual, as these F-pattern and F0 values were indeed taken from the same individual’s non-contemporaneous speech.

\[(4)\]

\[
\ln L_{\text{R, yes F-pattern, not too bad F0}} = (\text{score}_{\text{yes F-pattern}} \times \text{slope}_{\text{yes F-pattern}}) + (\text{score}_{\text{not too bad F0}} \times \text{slope}_{\text{not too bad F0}}) + \text{intercept}
\]

\[
= -2.03 \times 0.51 + 1.84 \times 0.4 + 2.66
\]

\[
= 2.36
\]

\[
= \text{LR 10.6}
\]

Although it is a well-established and powerful technique in automatic speaker recognition – see Rose (2012) for a demonstration of its power – fusion with multivariate logistic regression is not without its problems in a forensic context. Unlike the multivariate likelihood ratio, which takes the correlation between the actual variables into account, it deals with correlation between the scores arising from those variables. As Figure 15 shows, this is not the same thing. Figure 15 compares the Tippett plots for the same multivariate data fused with multivariate logistic regression and fused with the multivariate likelihood ratio. The data are the first three formants of the three vowels /o/ /ɛz/ and /æ/ from 153 speakers from Bernard’s (1967) database of broad general and cultivated Australian males. For the multivariate logistic regression fusion, within-segment scores were obtained with the multivariate likelihood ratio in the normal way for each vowel separately, and then fused. (The individual Cllrs for each of these vowels, from Rose (2010b: 32), were poor: 0.82 /o/, 0.9 /ɛz/,
0.72 /æ/). For the multivariate likelihood ratio fusion, scores were obtained with the multivariate likelihood ratio for all three vowels combined, and then calibrated with univariate logistic regression. It can be seen that the latter fusion, with the multivariate likelihood ratio, is considerably superior, with a $Cllr$ of 0.43 compared to 0.63 of the logistic regression fusion (note that both fusions improve on the results for the individual vowels, and are therefore successful). It is naïve to expect that both fusions would give the same result, as they use different methods (although logistic regression is also used with the multivariate fusion for calibration). This difference in performance clearly warrants further investigation, but in FVC logistic regression fusion currently remains the most straightforward option for combining the scores from different features into an overall LR, and the only one when features are not multivariately related.

Figure 15: Tippett plots comparing strength of evidence from logistic regression fusion and multivariate fusion of 153 speakers’ /ɔ/ /ɛː/ and /æ/ vowels.
The demonstrations just given of combining evidence from different features by fusing their scores with logistic regression are important not only because this would have allowed all the evidence quantified in the real case to have contributed to an outcome; as also made clear, it would have contributed to perhaps the most important consideration of all: being able to demonstrate the validity of the overall system. I think it is reasonable for the trier-of-fact to expect that we demonstrate that our forensic systems, or methods, do, in fact, work; that they are valid. As shown above, this can be done with Tippett plots, and probabilities of error.

In the real case, all I did was estimate the strength of evidence by evaluating the difference between the suspect and offender samples against a reference sample. Although it would have been possible to partially demonstrate the validity of each individual feature, it would not have been possible to do so for the overall system. As I have shown, such a demonstration, however, actually emerges as part of using logistic regression to fuse features. It would be useful to have terms for these two different methods. Since the former – the one actually used in the case – only makes use of the distributions of the features, it can be termed generative. The later development, making use of both distributions and logistic regression, can be termed generative-discriminative.

Unfortunately, I cannot demonstrate the generative-discriminative approach on the actual case data, as there is insufficient data in the reference sample to match the suspect and offender data, and these need to be matched for the logistic regression coefficients to be properly estimated (Morrison 2012b: 194–195). For example, to get regression coefficients for the yes F-pattern score one would require one sample to have 6 replicates and one 23, and the not too bad samples would have needed to be able to test 1 replicate against 15. The importance of using the correct numbers of replicates is shown in Figure 16. This shows the results of a Monte-Carlo simulation, using just the yes F-pattern data. Using the mean vectors and covariance matrices from the yes reference sample, multivariate LRIs were estimated permuting the number of suspect and offender replicates from 2 to 10 (offender) and from 2 to 28 (suspect). Thirty iterations were run for each permutation, from which the mean $Cllr$ was calculated and plotted. The resulting $Cllr$ surface in Figure 16 shows clearly that as the number of suspect and offender replicates is increased, the validity of the system as indicated by the $Cllr$ value improves. The improvement in validity from a system of the type just demonstrated, with just three suspect and offender replicates for yes F-pattern, to a system with 6 offender and 23 suspect replicates as used in the real case, is considerable: the $Cllr$ drops from 0.197 to 0.042.
Figure 16: $Cllr$ surface for Monte Carlo simulation of $yes$ F-pattern multivariate LR showing effect of suspect and offender replicate number on $Cllr$. Mean $Cllr$ values are indicated for a comparison with three replicates for both suspect and offender and for a comparison with 6 offender replicates and 23 suspect replicates.

Figure 17 shows the closest one can get to a generative-discriminative result with the $yes$ and $not too bad$ F-pattern and $not too bad$ F0 data used in the actual case. The Tippett plot is based on three replicates each for the $yes$ data, and one versus six replicates for the $not too bad$ data. As can be seen, even with much lower replicate numbers the validity is quite good, with a $Cllr$ of 0.063. All same-speaker comparisons were correctly discriminated, with 1.5% of different-speaker comparisons incorrectly evaluated. Bearing in mind the Monte Carlo results just demonstrated, it is likely that with the much higher replicate number of the real case, the validity would be considerably better than shown in Figure 17.

The effect of incorporating the more recent methods is an interesting and empirical one, and it will be instructive to see what kind of a difference the use of trajectory features will make. In particular, the relationship between the purely generative and the generative-discriminative approaches needs more research. At the time of the trial, however, to which we now turn, these improvements lay in the future.
Figure 17: Tippett plot for all five generative-discriminative fused yes and not too bad features, with replicate numbers of 3 vs 3 (yes) and 1 vs 6 (not too bad).

The LR goes to court

The final point in the FVC process was of course the trial \((R \, v \, Hufnagl, \, New \, South \, Wales \, District \, Court \, [2008])\). This is where the science ends and the law begins. In this section I describe my perception of how the LR-based evidence was received in court (one juror, who was judicially admonished for falling asleep, obviously didn’t quite receive as much as the others). I have also reflected on problems with LRs, arising from my perception.

Leading and cross-examination strategies

Prosecution and defence strategies to lead/cross my evidence had to be based on their assessment of what was best going to convince a jury, and this is almost certainly not going to be academic arguments relating to the strengths and weaknesses of the LR approach in general, or of the multivariate LR formula in particular. So there is no way of knowing to what extent the basics of the LR approach were actually understood. Prosecution seemed to put empha-
sis on my showing a redundantly large number of spectrograms, perhaps thereby trying to emphasise the idea to the jury that the approach was scientific. Defence suggested the LR approach was somehow my personal development, hoping to imply, I think, that it was not widely accepted. It is certainly true that the LR approach is not widely accepted, but not because I developed it (I didn’t: you have to go back a few centuries for the person responsible). Defence also argued that in selecting the reference sample I had not taken into account the fact that the suspect was from a particular area of Sydney. While it was good to see the question of the reference sample broached, this is a spurious argument. In asking what the probability is of getting the offender data assuming it had come from the suspect, one does indeed partially condition on the suspect. However, the reference sample is chosen with respect to a plausible defence hypothesis, and that must sensibly relate to the offender’s voice, not the suspect’s (and in any case the features used were not such as to be expected to vary across different areas of Sydney).

I felt that it was not wise to try to explain such things to a jury. (I expect that the reader, having waded through my attempts to explain the details of a LR approach in the preceding sections, will agree that this would have been far too technical, and pointless, and might well have made even more jurors comatose). Instead I tried to concentrate on emphasising two things. Firstly, that I was trying to estimate the strength of the evidence and not the probability that the suspect said the incriminating speech. Secondly, that the jury should not give much weight to the precise value of the LR; only that it was very big. It is encouraging to report that I felt tremendously aided in this by her Honour, who insisted that I repeat these ideas many times, so that the jury might have a chance of understanding them.

What is the court to do with a likelihood ratio?

The forensic expert is supposed to provide the trier-of-fact with an estimation of the strength of the evidence. That’s all very well, but what is the trier-of-fact supposed to do with it? A well-known probability expert has cautioned: ‘one must be wary of over-simplistic direct interpretation of the numerical value of the likelihood ratio, which can only be sensibly considered in conjunction with other information’ (David 2002: 130). This other information is the prior odds, and ideally the trier-of-fact is supposed to combine the prior odds and the LR to arrive at the posterior odds (whence the probability) of the hypothesis, given the evidence.

Presumably it is the expert’s job to explain the meaning of the LR, but forensic statistics textbooks, for example the chapter on explaining the strength
of evidence in Robertson and Vignaux (1995), do not give much explicit help on how this should happen. Likelihood ratios have been a part of DNA evidence in court for some time, so perhaps there is help from that quarter? No: Balding (2005: 145), for example, writes:

If any … reader was hoping that I would be able to prescribe a formulaic approach to reporting DNA evidence in court, that satisfies the needs of jurors and the demands of judges in every case, then s/he will be disappointed. I have no magic formula to overcome the difficult issues that arise in presenting complex scientific evidence to non-expert judges and juries.

From a theoretical point of view I see this as a major problem with the use of a LR-based approach, since the magnitude of the evidential LR has no meaning outside its matrix theorem. For example, with prior odds of 1 to 100 against – there are 101 other people including the suspect who could have said the incriminating speech – a LR of 1000 means a posterior probability of c. 91% in favour of them saying it; but with priors of 1 to 1000 – 1001 people including the suspect who could have said it – the same LR gives a posterior of only 50%, meaning you logically cannot say, on the basis of the evidence, whether they did or not. This means that proposals to explain LRs verbally in terms of varying degrees of strength of evidence in support of the defence or prosecution hypotheses are misleading: a LR described as indicating a high degree of support for the prosecution can be transformed into a posterior supporting the defence by sufficiently large priors, and vice-versa. Perhaps a more effective strategy – if sufficient confidence exists in the actual value of the LR – is to explain it in the way I have tried above, namely in terms of the number of others necessary for a given posterior. To say, for example, that if you want to be n% confident that the suspect said the incriminating speech, other evidence would have to have narrowed down the other possible speakers to no more than m.

**Combining quantitative and qualitative evidence**

In the reality of a trial there remains another big problem with trying to make sense of strength of evidence estimates in terms of Bayes’ Theorem (quite apart from the fact that the way it works is not well understood by the legal advocates whose job it would be to explain it). This problem was pointed out in 2002 by Justice Hodgson – a rare combination of appeals court judge and probability expert. Also clearly in favour of LRs for quantifiable forensic evidence (Hodgson 2007), he is thus worth heeding.

The point is simply that since not all types of evidence in a trial can be
sensibly assigned a LR, there is no way of mathematically combining à la Bayes the LR-based evidence with the non-numerically based evidence. The Law expects the jury, after all, to evaluate the evidence using their commonsense. As Hodgson also points out, this is one reason why another popular attempt to explain the meaning of the LR to the jury – e.g. *whatever your belief in the hypothesis was before the evidence you must increase/decrease it by the amount of the LR* – is also not going to work.

In the trial, no attempt was made to explain to the jury the meaning of the LR-based voice evidence: it was not led, it was not crossed; and all the jury had was an unqualified statement that they were much more likely to get the differences between the suspect and offender speech samples if they had come from the same person rather than from different people. Perhaps, given the problems associated with explaining the idea of strength of evidence, it was better so.

We thus end up, at least in LR-based FVC, with a current impasse at the boundary between science and the law. But it is an impasse, apparently, that is unsatisfactory and frustrating only from the point of view of the scientist. I have no idea of course of what sense, if any, the jury made of my LR-based voice evidence. Whichever way it was construed, and combined with the other evidence, they returned a verdict of guilty (ABC 2008, CDPP 2009–2010).

**Summary and conclusion**

I have tried in this article to give an idea of what is involved in estimating a likelihood ratio using data from a real case (the *science* part of the article’s title), and what happens when it is given to the trier-of-fact (the *law* part). I have shown how, apart from conforming to the requirements of logic, and thereby making it necessary for the expert to address impartially the probability of the evidence under both defence and prosecution hypotheses, the method can arguably arrogate the two watchwords of transparency and testability. It can be made clear to the point of replicability what was done, and the considerations motivating analytical decisions and inferences at every stage. This is important, should the claim be challenged on any assumption. For example, if it is claimed by defence that the reference sample is defective, a proposed replacement can be tested to see if it substantially affects the outcome – the defence would hope that the value of the LR would be reduced to near unity, or even substantially reversed. The same can be done if it is claimed that the selection of features was defective by omitting features vital to the outcome. Perhaps the VOT of *too* should have been included? Well, test it and see! The testability of the approach is its most important aspect: I have shown how the approach permits, and
indeed in its later generative-discriminative developments requires, a demonstration of how well same-speaker samples can be discriminated from different-speaker samples under conditions reflecting the actual case. It would be nice if the trier-of-fact actually understood enough to demand this of their experts.

One aspect of the approach worth mentioning that will not have emerged clearly enough from looking at a single case is that there is no one-size-fits-all. Real-world cases are never the same, and the details have to be chosen to fit the case. This case was unusual in that its small amount of offender data contained material useful for estimating a LR. Because of the small amount of offender data, and the mismatch between the offender and some of the suspect recordings, it is unlikely that an automatic system could have been recruited to act as a base-line system in the way demonstrated, for example, in Zhang et al. (2013). Although phonological knowledge is always implicit in the acoustic-phonetic approach to LR estimation – we would otherwise not know which vowel allophones, if any, are comparable – the case is perhaps a nice demonstration of how knowledge of intonational phonology can be of use. But no one-size-fits-all does not mean no one-approach-fits-all: there is certainly no other way of rationally assessing the value of the evidence than to say how likely it is under competing defence and prosecution hypotheses.

The theoretically least satisfactory aspect of the forensic voice comparison, I felt, was unfortunately, but perhaps unavoidably, its most important: reception of the voice evidence in court. I mentioned two of the main problems with LR quantified evidence: explaining its meaning to the trier of fact with reference to prior odds, and its proper integration with the other evidence, both of which tend to receive short shrift in forensic statistic textbooks, and deserve more attention.

But you have to understand before you can explain. Mlodinow (2008: 40) observed:

statistical arguments are becoming increasingly important because of the necessity of assessing DNA evidence. Unfortunately, with this increased importance has not come increased understanding on the part of attorneys, judges, or juries.

I don’t think it’s fair to blame the juries. Given that the Law still obviously has problems with understanding the rational evaluation of the type of forensic evidence presented here, and given that our main purpose is to serve the Law, perhaps we should accommodate to it and set our sights lower. Perhaps the courts would be happier just with an indication of the degree of similarity between suspect and offender speech samples, presented along the lines of
Figures 2, 3, 4 and 8? On the other hand, the accused has a right to a fair trial, and this is unlikely to occur if the principles of rational evaluation of forensic evidence are poorly understood by defence. How would you then rebut a prosecution demonstration that the suspect and offender samples were very similar? So perhaps now amelioration really does have also to come from a better understanding of the approach from the legal profession.

About the author

Phil Rose was Reader at the Australian National University, where he taught Phonetics and Chinese Linguistics for 30 years. He has been a British Academy visiting professor at the University of Edinburgh’s Joseph Bell Centre for Forensic Statistics and Legal Reasoning, is chairman of the Forensic Speech Science Committee of the Australasian Speech Science and Technology Association and a member of the Australian Academy of Forensic Sciences. He has done forensic voice comparison case-work on Australian English and varieties of Chinese for nearly 20 years.

Acknowledgments

I thank my reviewers for taking their time to make many useful suggestions, most of which I have been able to incorporate. Their questions made it clearer to me which parts needed more explanation and I hope I have rectified that within the limits of what was already a very long, dense article. Thanks also to my long-suffering reference sample speakers, who, like me, can no longer say not too bad without shuddering inwardly.

Notes

1 This is an expanded version of my keynote at the 7th Iberspeech Speech Technology Conference 2012 and my paper at the 14th Australian International Conference on Speech Science and Technology 2012.

2 ‘Steam-engine time’ is a metaphor for the widespread acceptance of an idea or invention which, like the steam engine, has made a sudden polygenetic emergence after actually being around for a long time (Pratchett et al. 2005: 236–237, 243–244).

3 The term pitch is commonly found referring to both the perceptual property of the linguistic category of intonation and its acoustical reflex of F0. There are examples in this article of utterances having the same F0 but different intonational pitch (compare the F0 values and tone on not in Figure 8), and therefore I use pitch only in its proper, perceptual, sense.
I am making use here of Nolan’s (1983) characterisation of the information content of a voice arising from the interaction of choice and constraint. This model is by far the most useful in forensic voice comparison because it draws attention to and clarifies the sources of both between-speaker variation and within-speaker variation in the speech signal.

The use of credible intervals to quantify precision is contentious, for statistico-ontological reasons (Balding 2005: 92–93). It is not clear whether a ‘true’ LR exists such that it can be said to lie within the intervals at a certain probability; the uncertainty a credible interval is supposed to quantify is already embodied by a LR; and the probabilistic nature of the credible interval is too easily construed as just that – an interval, when it is actually a posterior distribution. Such arcane considerations are unlikely to trouble a court for the foreseeable future, whereas it is likely to welcome the pragmatic proposal, however contentious, of the expert being able to set an upper and lower bound on their LR estimate.

posterior odds = prior odds * LR. 1:1 (prior odds) * 10.6:1 (LR) = 10.6:1 (posterior odds) = 10.6/(10.6 + 1) = 91% posterior probability.

There are two references to R v Hufnagl in the NSWDC case reports for 2008 (www.austlii.edu.au/au/cases/nsw/NSWDC/2008/). Neither refers to the trial described in this paper. The trial described in this paper is not reported, as no point of law arose from it.

References

ABC (2008) http://www.abc.net.au/pm/content/2008/s2451596.htm


http://dx.doi.org/10.1016/j.csl.2005.08.001


http://dx.doi.org/10.1111/j.1751-5823.2003.tb00207.x

http://dx.doi.org/10.1558/ijsll.v18i2.293

http://dx.doi.org/10.1016/j.csl.2005.08.005


http://dx.doi.org/10.1016/j.forsciint.2011.12.009


