Where the Science Ends and the Law Begins: Theory and Reality in Likelihood Ratio-based Forensic Voice Comparison.

Phil Rose

Hong Kong University of Science & Technology and Australian National University
philip.rose@anu.edu.au

Abstract. The first use of Likelihood Ratios for the evaluation of acoustic-phonetic forensic voice comparison evidence in a real trial in Australia is described and critiqued.

Keywords. Forensic voice comparison, likelihood ratio, F-pattern, fundamental frequency.

1 Introduction

In forensic voice comparison (FVC), speech recordings from an unknown voice, usually of an offender, are compared with recordings from a known voice, usually the suspect. The aim is to help the trier-of-fact (in the case described here, a 12 person jury) decide whether the suspect has said the incriminating speech. Currently FVC, as reflected in research and practice, can be divided into two types depending on what the expert considers to be their ultimate purpose. In the first type, the expert considers their aim to be to say how likely it is, given the evidence, that the suspect said the incriminating speech. For example in a 2004 high profile murder case one of the leading UK forensic speech scientists was described as saying "I came to the view that there were very strong similarities and no differences of a kind that could cause me to eliminate [the suspect]." and "I would consider the likelihood of the voices coming from two different people to be remote." [1]. 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) – in other words, to estimate 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.

For some time now, the use of a LR has been theoretically recognised as the correct logical framework for the evaluation of forensic evidence. It is referred to in all major text-books on forensic statistics, e.g [2 – 3], and has been endorsed by several forensic institutions, e.g the UK and Irish Association of Forensic Science Providers has made it part of its *Standards for the formulation of evaluative forensic science*

expert opinion [4]. It is logically correct, since by Bayes' Theorem a posterior probability – like "it is highly likely the suspect said the incriminating speech" – cannot be estimated absent prior odds, to which the expert is not usually privy. This is why the first type of FVC, unless the prior odds are 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 [5 – 6].

Apart from its correctness, the LR approach 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, it also allows an expert in an adversarial system to be 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 tested [7].

As far as practice is concerned, a LR approach has also been implemented as a matter of course since the mid 1990's in DNA profiling [8] (although it is not clear that the courts actually understand that the random occurrence probability usually quoted is part of a LR). 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 [9]. 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. [10], and it has been shown in [11] 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" [3]. The LR-based testing of other forensic evidence types is now following: fingerprints [12], handwriting [13] and SMS texting [14].

Despite these encouraging developments in FVC, which have been interpreted as an incipient paradigm shift towards a maximally objective evaluation of forensic evidence [9], 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 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 recently maintaining that "Interpretation [of evidence] includes answering the question as to whether or not …items share a common origin…" [15]. 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 aforementioned Australian draft standards proposed a probability-of-hypothesis-given-evidence approach while simultaneously recommending a text-book on the interpretation of evidence [5] which clearly espouses the opposite, probability-of-evidence-under-competing-hypotheses likelihood-ratio approach. The degree of misunderstanding by the courts is probably best exemplified by the mind-boggling confusion and inconsistencies, described in detail in [16], in a 2010 ruling by the England and Wales Court of Criminal Appeal in *R v T* [17].

It is not just the legal profession that finds the approach difficult to understand, however. A recent book on Forensic Linguistic evidence misrepresents LRs [18], and perhaps the best-known phonetics text-book [19] 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 *R v T* ruling [16]). Steam-engine time[1] for the likelihood ratio approach in FVC still looks a very long way off.

I started to use LR estimates in my case-work in 2002 (some details from actual cases are in [21 – 22]), 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" [23], one aim of this paper is therefore to briefly document the first FVC case where LR-based speech evidence was received in court in Australia, and to try thereby to give an idea of what is involved. I will also briefly address areas where the LR estimation might have been substantially improved, given what we know from research in the interim. A third problem is also broached, concerning the problems of conveying the meaning of LRs to a jury, and integrating any LR-based evidence with the other evidence in a case.

2 How to make $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 call-back on the fax (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 = J.P. Morgan employee, F = fraudster):

E *J.P. Morgan Greg speaking*
F **yeah hello Greg this is Craig Slater here mate**

[1] '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 [20].

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 $150 million short.

3 Approach

When using traditional features to estimate a LR, one first scours the offender and suspect data for comparable material, e.g. same utterance; same phonemes in comparable environments. 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. There were, however, 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. (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 his mates. The bulk of the LR estimate in this case thus rested on features extracted from *yes* and *not too bad*.

In order to estimate the probability of getting the difference between the offender sample features and another person chosen at random from the relevant population (the LR denominator) a reference sample is of course necessary. In this case, no explicit alternative hypothesis having been nominated by defence, it was sensible to assume that the offender voice belonged to a male speaker of Australian English, and I amassed a sample of 35 adult males with ages uniformly distributed between early 20's and early 70's who agreed to be phoned up and recorded. Unlike, for example, *I'll blow yer fuckin' head off*, the test material in this case has the enormous benefit of easy ecologically valid elicitation: reference sample speakers were instructed to appropriately reply *not too bad* and *yes* or *no* to a set of questions I asked, like *how's it going*? or *are you inside*? I tried to indirectly prime reference subjects 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. In this way it was possible to obtain several replicates of *yes*, and *not too bad* with the correct 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.

4 F0 in *not too bad*

Although parameters of long term F0 distributions have been shown to perform rather well in LR-based discrimination research [24], F0 is usually not of much use in real case-work because of its disadvantageous short-term variance ratio. This case was different, however, in that all test data *not too bad*s are said in response to the interlocutor's asking how the suspect/offender is. They have very similar intonational structure typical for this conversational interchange, conveying the typical rise nuclear tone meanings of supportive interest encouraging further conversation [25]. 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[1] (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 means that the F0 values on *not too bad* are highly comparable and might be expected to yield useful likelihood ratios (i.e. values that deviate substantially from unity). Figure 1 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 ca. 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 2 compares the offender F0 with the F0 of the suspect's 15 *not too bad* utterances. The similarity is considerable, with the offender's F0 time-course lying completely

[1] The term *pitch* is commonly found referring to both the perceptual property of the linguistic category of intonation and its acoustical reflex of F0. In this paper it has the former reference only.

within, and in some places almost exactly in the middle of, the suspect's distribution. Note too the suspect's use of both H and L on *not*.

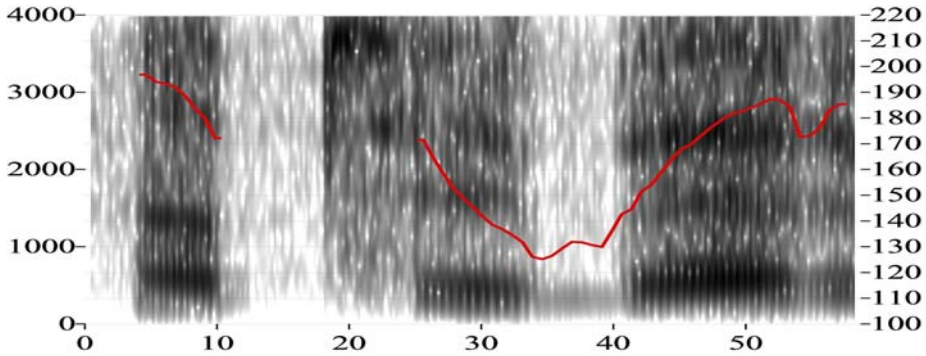

Fig. 1. F0 time course of offender's *not too bad* (right-hand scale in Hz) superimposed on wide-band spectrogram (left-hand scale in Hz). x axis = duration (csec.).

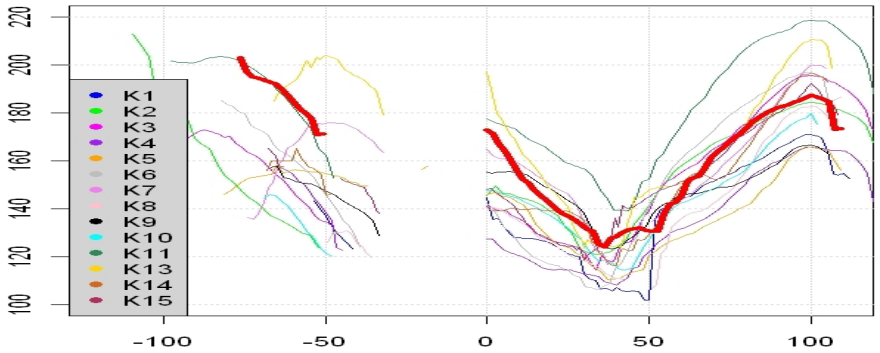

Fig. 2. Suspect (K) and offender (red) *not too bad* F0 plotted as function of equalized duration. Axes: vertical = F0 (Hz), horizontal = duration equalized around /u:/ onset (0%) and peak F0 in bad (100%).

The *not too bad* F0 values from the reference sample speakers showed no surprises in variance structure. Figure 3, 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 3 displays this behaviour: in recording sessions 1 & 2 the different F0 height on his *not* corresponding to the H vs. L distinction can be easily seen. This speaker also shows considerable non-phonological variation in F0 between sessions: his *not too bad*s in session 2 are much higher and sound more enthusiastic that in session 1, for example. A comparison with the second speaker in figure 3 illustrates typical aspects of between-speaker variation.

Most obviously, he shows 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 session to session.

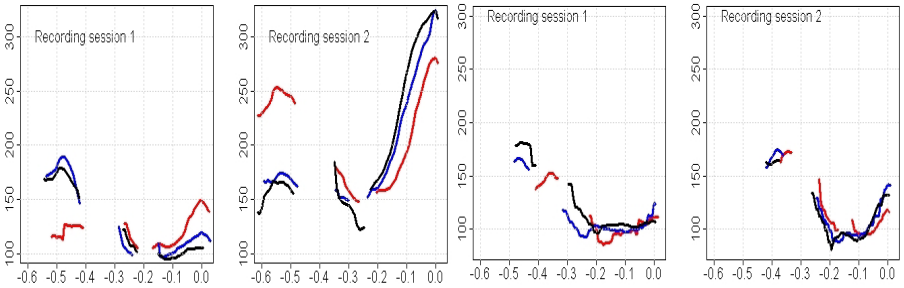

Fig. 3. Within- and between-speaker variation in reference sample: *not too bad* F0 tokens from two speakers on two different occasions. Axes: vertical = F0 (Hz), horizontal = duration (csec. from peak F0 in bad).

The F0 time course in *too bad* was sampled at four points: mid point of the vowel in too, and at first target, peak and mid way between in the vowel in *bad*. Because of the aforementioned associated phonological variation between H and L, F0 on not was not used. The four sampled F0 values were treated as multivariate data and LRs were estimated using a generative multivariate LR formula [26], modeling 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 or kernel modeling of the reference sample. Given the high degree of similarity between suspect and offender samples (figure 2), the relatively low LR value for the comparison is salutary and reminds us that evidentiary value is also dependent on typicality.

5 Formant pattern in *yes*

Both suspect and offender *yes* tokens sounded to have a more centralized /e/ than normal, and lower than normal pitch on /s/ (somewhat reminiscent of the pharyngealised /s/ in Arabic). Figure 4 shows spectrograms with superimposed formants of two *yes* tokens from offender and suspect. Both show unremarkable F-pattern configurations given the auditory impression – note the lower than normal cut-off frequencies for the /s/ (four formants were extracted below 3.6 kHz, so the F-pattern of the /s/ is not tracked well). The offender and suspect /je/ F-pattern was sampled at onset, midpoint and offset. Figure 5 shows 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. Thus the probability of getting the offender values assuming they have come from the suspect will be fairly high. Figure 5 also shows the distribution of the reference sample values for /je/. It can be seen that the between-speaker

variance in F2 and F3 is quite big, but more importantly it shows that the F-pattern in the suspect and offender has unusually low offset values. This presumably correlates with the audibly lower, more centralized nucleus. This means that a LR greater than unity is to be expected. A MVLR was estimated for the F2 and F3 vales at all three sampling points in /je/, as it was assumed that the F1 would have been compromised by the telephone transmission. Both normal and kernel density MVLRs indicated that the difference between suspect and offender /je/ F-pattern was about 70 times more likely had they come from the same rather than different speakers.

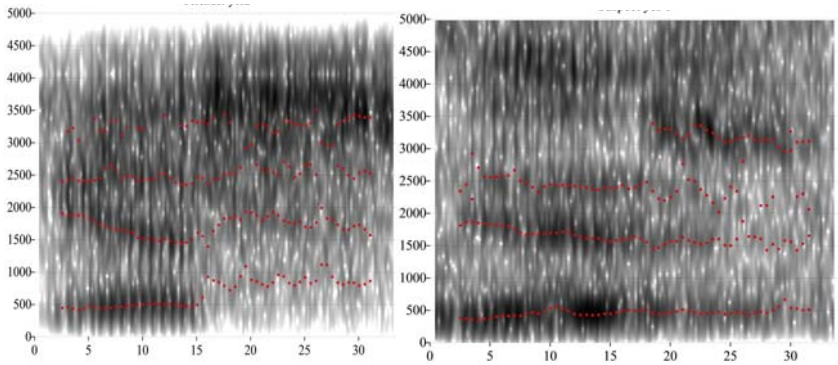

Fig. 4. Spectrograms of offender (left) and suspect (right) *yes* tokens. x axis = duration (csec.)

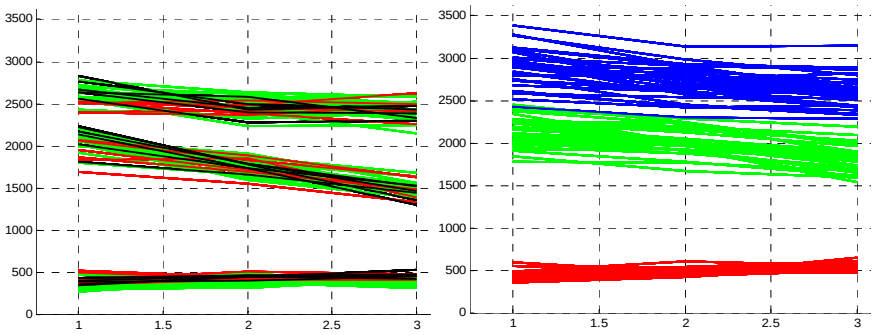

Fig. 5. /je/ F-pattern values in *yes* sampled at onset, midpoint and offset. Left: individual tokens from offender (black) compared with suspect (red, green). Right: mean values from reference sample. Axes: vertical = frequency (Hz), horizontal = equalized duration.

6 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*, /u:/ in *too* and /æ/ in *bad*) were also estimated. As expected, comparison with /o/ was complicated due

to nasal poles associated with quasi-Helmholtz and nasal cavity resonances induced by /n/. Figure 6 compares the F-pattern of offender and suspect in these three vowels (nasal poles in /o/ are not shown). 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. The F-pattern differences were parametrised by mean values taken over quasi-steady-state portions at vocalic target, and point measurements at offset. Figure 7 shows some of these values against the corresponding reference sample distributions, where it can be seen that they are fairly typical (as in *yes*, some values were discarded because of possible influence from the telephone transmission). The resulting MVLR estimates were all greater than unity: ca. 24, 5 & 11 for /o u: æ/ respectively. In addition, I attempted a crude (because discrete) LR estimate for the low cut-off in the /s/ spectrum in *yes*: this showed that a low cut-off was conservatively about 2.5 times more likely in both samples if they had come from the same rather than different speakers.

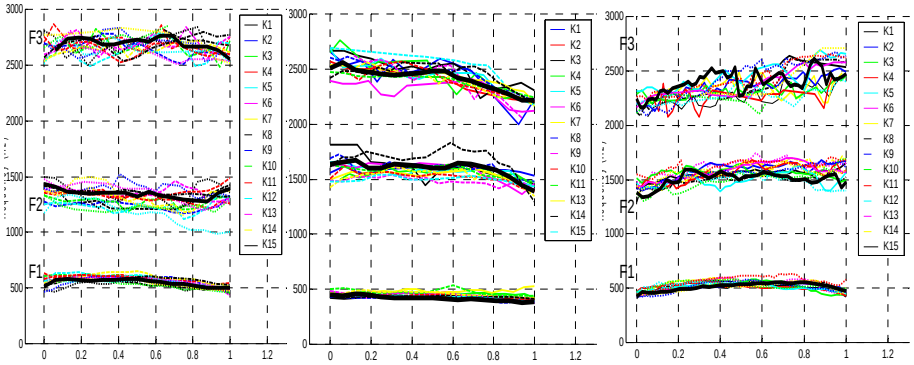

Fig. 6. Offender's F-pattern (thick line) plotted against suspect's values in the vowels of *not* (left) *too* (middle) and *bad*. Vertical axis = frequency (Hz), horizontal axis = equalized duration (% * .01).

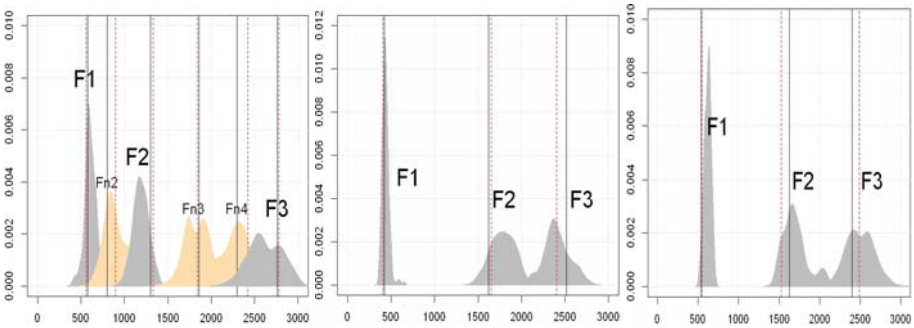

Fig. 7. Target F-pattern means for offender (vertical dotted red lines) and suspect (vertical solid black lines) in the vowels in *not* (left), *too* (middle) and *bad* plotted against the reference sample distributions. Fn = nasal pole. Vertical axis = probability density, horizontal axis = frequency (Hz).

7 Result

A naïve Bayes combination of the LRs from the features described yielded an exceptionally large overall LR of about 11 million (one feature not described contributed an additional LR of ca. 2). Since I suspected some between-segment correlation, but could not estimate it, I simply discarded the putatively correlated LRs (e.g. from individual formants in *not too bad*) to arrive at a much smaller LR estimate of ca. 300,000. According to Bayes' Theorem there would now have to be more than roughly 16,000 others who could have said the incriminating speech before the posterior probability that the suspect said it fell below 95%.

8 Critique

The now five years since this case have seen continuing improvements in traditional FVC LR estimation. The main ones have been in the recruitment of various aspects of automatic FSR approaches to (1) better parametrise traditional features like formants [27]; (2) combine LRs from different features with logistic-regression-fusion, thus providing a potential solution to the problem of possible between-segment correlation [28]; (3) use features like cepstral coefficients as well as formants to characterize segments [29], and (4) provide measures of accuracy and precision [7]. All of these 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 DCT or polynomial coefficients, rather than use point estimates. Secondly, the problems with handling possible between-segment correlation might have been obviated by logistic-regression fusion of the LRs from the different segments. Finally, the whole of the /s/ spectrum in *yes* could have been compared using cepstrally-mean-subtracted CCs, rather than just the fact that it had a low cut-off. The effect of incorporating these methods is an interesting and empirical one, and I am currently re-estimating the LR for the evidence with them to see what sort of a difference they make. At the time of the trial, however, to which we now turn, these improvements lay in the future.

9 Reception

The final point in the FVC process was of course the trial. This is where the science ends and the law begins. My perception of how the LR-based evidence was received in court is as follows (one juror, who received judicial admonishment for continuously falling asleep, obviously didn't receive much). Prosecution and defence strategies to lead/attack 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. Prosecution

seemed to put emphasis 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, and by implication therefore not widely accepted. They 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. 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 the 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. Instead I tried to concentrate on emphasizing two points. 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 the judge, who insisted that I repeat these ideas many times, so that the jury might have a chance of understanding them.

A well-known probability expert has cautioned: "… one must be wary of oversimplistic direct interpretation of the numerical value of the likelihood ratio, which can only be sensibly considered in conjunction with other information"[30]. However, no attempt was made in the trial to explain to the jury the meaning of the LR-based voice evidence by introducing this extra information – the notion of combining it with prior odds to estimate a posterior. From a theoretical point of view, of course, this is a major problem with the use of a LR-based approach, since the magnitude of the evidential LR has no meaning outside of its matrix theorem (e.g. with prior odds of 1 to 100 against, a LR of 1000 means a posterior probability of ca. 91% in favour; but with priors of 1 to 1000, the same LR gives a posterior of only 50%). 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 *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*. In the reality of a trial, however, there remain big problems 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 advocates whose job it would be to explain it). The main problem was pointed out by Justice Hodgson [31] – a rare combination of appeals court judge and probability expert, and thus worth heeding. It is simply that since not all types of evidence in a trial can be sensibly assigned a LR there is no way of 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.

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 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 [32 – 33].

10 References

1. http://news.bbc.co.uk/2/hi/uk_news/england/west_yorkshire/4039057.stm
2. Aitken, C.G.G., Taroni, F.: Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, Chichester (2004)
3. Balding, D.J.: Weight of Evidence for Forensic DNA Profiles. Wiley, Chichester (2005)
4. Association of Forensic Science Providers: Standards for the formulation of evaluative forensic science expert opinion. Science & Justice 49, 161 – 164 (2009)
5. Robertson, B, Vignaux, G.A.: Interpreting Evidence – Evaluating Forensic Science in the Courtroom. Wiley, Chichester (1995)
6. Morrison, G.S.: Forensic Voice Comparison. In: Freckleton, I., Selby H. (eds.) Expert Evidence. Ch.99 (2010)
7. Morrison, G.S.: Measuring the Validity and Reliability of forensic likelihood-ratio systems. Science & Justice (51) 3, 91-98 (2011)
8. Forman, L.A., Champod, C., Evett, I.W., Lambert, J.A., Pope, S.: Interpreting DNA evidence: a review. International Statistics Journal 71 473 – 495 (2003)
9. Morrison: G.S. Forensic voice comparison and the paradigm shift. Science & Justice 49, 298–308 (2009)
10. Gonzalez-Rodriguez J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech and Language 20, 331 – 355 (2006)
11. Gonzalez-Rodriguez J., Rose P., Ramos, D., Torre, D. & Ortega-García, J.: Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. IEEE Trans. on Audio Speech and Language Processing 15(7), 2104 – 2115 (2007)
12. Neumann, C., Evett, I.W., Skerrett, J. Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. Journal of the Royal Statistical Society 175, 371 – 415 (2012)
13. Hepler, A.B., Saunders, C.P., Davis, L.J., Buscaglia, J.: Score-based likelihood ratios for handwriting evidence. Forensic Science International, (2012)
14. Ishihara, S.: A Forensic Text Comparison in SMS Messages: A Likelihood Ratio Approach with Lexical Features. In: Clarke, N., Tryfonas, T., Dodge, R. (eds.) Proc. Seventh Intl. Workshop on Digital Forensics and Incident Analysis, 55 – 65 (2012)
15. Standards Australia: Forensic Analysis Part 3: Interpretation. Draft for Public Comment, DR AS 5388.3 http://infostore.saiglobal.com/store/Details.aspx?ProductID=1530224 (2012)
16. Morrison, G.S.: The Likelihood Ratio framework and forensic evidence in court: A response to *R v T*. International Journal of Evidence and Proof 16, 1 – 29 (2012)
17. EWCA 2439 (2010)
18. Morrison, G.S.: Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. Australian Journal of Forensic Sciences 41, 155 – 161 (2009)

19. Ladefoged, P: A Course in Phonetics. 5th ed. Thomson, Boston (2006)
20. Pratchett, T., Stewart, I, Cohen, J.: Darwin's Watch. Ebury Press, London (2005)
21. Rose, P.: The Technical Comparison of Forensic Voice Samples. In: Freckleton, I., Selby H. (eds.) Expert Evidence. Ch.99 (2002)
22. Rose, P.: Technical forensic speaker recognition: Evaluation, types and testing of evidence. Computer Speech and Language 20, 159 – 191 (2006)
23. Berger, C.: Criminalistics is reasoning backwards. Nederlands Juristenblad AFL13 (2010)
24. Kinoshita, Y., Ishihara, S., Rose P.: Exploring the Discriminatory Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition. Intl. J. Speech Language & the Law 16(1), 91 – 111 (2008)
25. Wells, J. C.: English Intonation. Cambridge University Press, Cambridge UK (2006)
26. Aitken, C.G.G., Lucy, D.: Evaluation of trace evidence in the form of multivariate data. Applied Statistics 53(4), 109 – 122 (2004)
27. Morrison, G.S.: Likelihood Ratio forensic voice comparison using parametric representation of the formant trajectories of diphthongs. JASA 125, 2387 – 2397 (2009)
28. Pigeon, S., Druyts, P., Verlinde.: Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. Digital Signal Processing (10) 1-3, 237 – 248 (2000)
29. Rose, P.: Forensic Voice Comparison with Secular Shibboleths – a hybrid fused GMM-Multivariate Likelihood Ratio-based Approach Using Alveolo-Palatal Fricative Cepstral Spectra. Proc. ICASSP, 2011.
30. David, P.: Statistics and the Law. In: Bell, A., Swenson-Wright, J., Tybjerg, K. (eds.) Evidence. Cambridge University Press, Cambridge UK, 119 – 148 (2002)
31. Hodgson, D.: A lawyer looks at Bayes' theorem. The Australian Law Journal 76, 109 – 118 (2002)
32. http://www.abc.net.au/pm/content/2008/s2451596.htm
33. Commonwealth Director of Public Prosecutions Report 2009-2010.