Yes, not too bad – Likelihood Ratio-Based Forensic Voice Comparison in a $150 Million Telephone Fraud

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 traditional forensic voice comparison evidence in a real trial in Australia is described and critiqued.

Index Terms: Forensic Voice Comparison, Likelihood Ratio

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, of course, 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 starkly 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. In the second, the expert's aim is to estimate 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 both theoretically recognised as the correct logical framework for the evaluation of forensic evidence, and implemented as a matter of course in some areas, e.g. DNA profiling. 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 cannot be logically correct. Since a posterior may well impinge on considerations of ultimate issue, the use of a LR may also be the legally correct option.

The use of LRs in FVC was an idea whose time came around the beginning of the new century, when its efficacy first began to be demonstrated both with automatic and traditional phonetic features [1]. The results from now well over a decade's extensive, and continuing, research testing have shown that the approach works rather well – same-speaker speech samples can for example be rather well discriminated from different-speaker speech samples on the basis of their LRs – and it has been shown that the approach can emulate the DNA gold-standard [2]. Despite this, it is safe to say the idea has not yet caught on. This is partly because of the inherent conservatism of the legal profession, and partly because, it appears, many of its practitioners find it difficult to understand. For example, 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…" [3]. Not just the legal profession finds it difficult, however. A recent book on Forensic Linguistic evidence misrepresents LRs [4], and perhaps the best-known phonetics text-book [5] 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.

I started to use LR estimates in my case-work in 2002 (some details from actual cases are in [6]), 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" [7], the aim of this paper is therefore to 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 how the approach works. I will also briefly address areas where the LR estimation might have been substantially improved, given what we know from research in the interim.

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**
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 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 [8], 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 bads* 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. 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 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 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*.

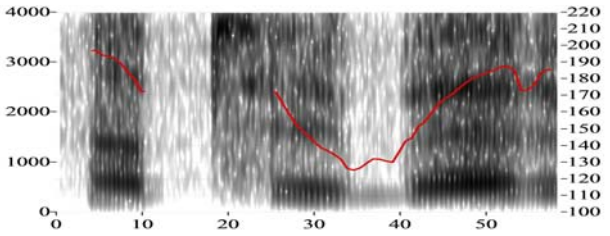

Figure 1: *F0 time course of offender's* not too bad *(right-hand scale in Hz) superimposed on wide-band spectrogram (left-hand scale in Hz). x axis = duration (csec.).*

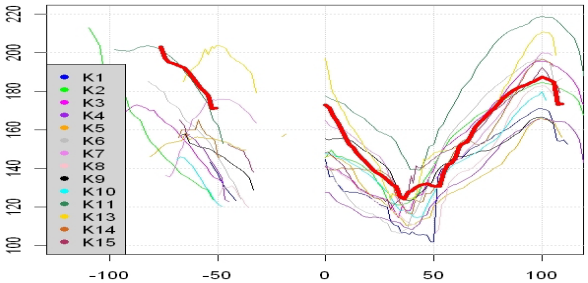

Figure 2: *Suspect (*K*) and offender (red)* not too bad *F0 plotted as function of equalized duration. Axes: vertical = F0 (Hz), horizontal = duration equalized around* /u:/ *onset (0%) and peak F0 in* bad *(100%).*

The *not too bad* F0 values from the reference sample speakers showed no surprises in variance structure. Figure 3, which plots the F0 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 bads* 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.

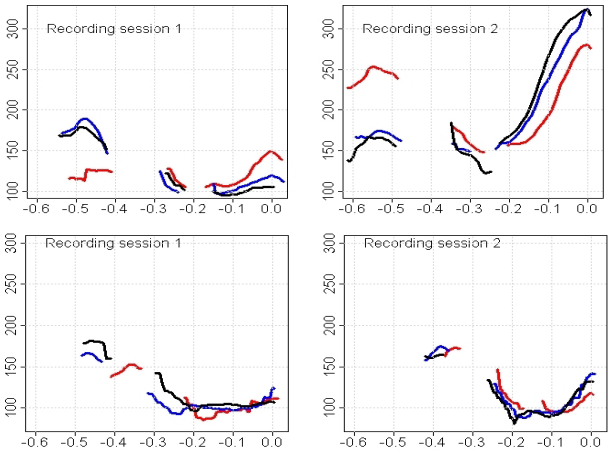

Figure 3: *Within- & between-speaker variation in reference sample:* not too bad *F0 tokens from two speakers on two different occasions. Axes: vertical = F0 (Hz), horizontal = duration (csec. from peak F0 in* bad*). Red = 1st, blue = 2nd, black = 3rd replicate.*

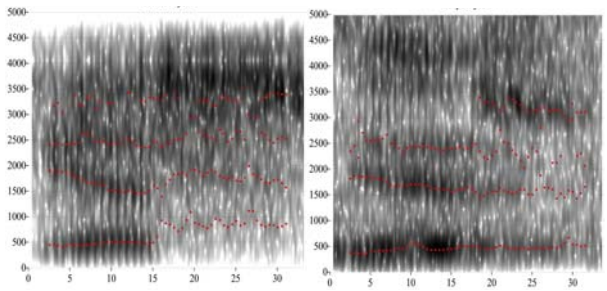

Figure 4: *Spectrograms of offender (left) and suspect (right)* yes *tokens. x axis = duration (csec.)*

4.1. Processing: LR for *not too bad* F0

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 [9], modeling the reference sample both normally and with a kernel density. The difference between the suspect and offender F0 values in *not 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.

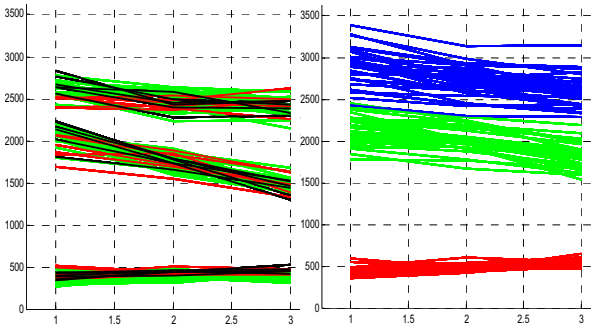

Figure 5: /je/ *F-pattern values in* yes *sampled at onset, midpoint and offset. Left: individual tokens from offender (black) compared with suspect (red, green). Right: mean values from reference sample. Axes: vertical = frequency (Hz), horizontal = equalized duration.*

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 /s̴/ 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, mid-point 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. This showed 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.

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. There is no space here to detail the results, other than to say that the comparison with /o/ was complicated due to nasal poles induced by /n/, one of which, however, was able to be included. The resulting LR 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.

7. Result

A naïve Bayes combination of the LRs from the features described yielded a very big 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 ca. 16000 others who could have said the incriminating speech before the posterior probability that the suspect said it fell below 95%).

8. 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 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 very very 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.

9. Critique

The now five years since this case have seen continuing improvements in traditional FVC LR estimation. The main improvements have been in the recruitment of automatic FSR approaches to (1) better parametrise traditional features like formants [10]; (2) combine LRs from different features with logistic-regression-fusion, thus providing a solution to the problem of possible between-segment correlation [11]; (3) use features like cepstral coefficients as well as formants to characterize segments [12], and (4) provide measures of accuracy and precision [13]. 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 would have been obviated by logistic-regression fusion of the LRs from the different segments. Finally, the whole of the /s/ spectrum could have been compared using cepstrally-mean-subtracted CCs, rather than just the fact that it had a low cut-off. I am currently estimating the LR of the evidence with these approaches to see what sort of a difference they make.

At the time of the trial, however, these improvements lay in the future. I have no idea, of course, of what sense the jury made of my LR-based voice evidence (they were given no theorem to estimate the posterior probability). Whichever way it was construed, and combined with the other evidence, they returned a verdict of guilty [14, 15].

10. References

[1] Morrison, G.S. "Forensic voice comparison and the paradigm shift", *Science & Justice*, 49: 298–308, 2009.

[2] 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.

[3] 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.

[4] Morrison, G.S. "Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework", *Australian Journal of Forensic Sciences*, 41: 155-161, 2009.

[5] Ladefoged, P. *A Course in Phonetics*, 5th ed. Thomson, Boston, 2006.

[6] Rose, P. The Technical Comparison of Forensic Voice Samples, in Freckleton & Selby (Eds.) *Expert Evidence*, 2002.

[7] Berger, C. "Criminalistics is reasoning backwards", Nederlands Juristenblad, AFL13, 2010.

[8] 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*, 2008".

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

[10] Morrison, G.S "Likelihood Ratio forensic voice comparison using parametric representation of the formant trajectories of diphthongs", *JASA,* 125, 2387-2397, 2009.

[11] 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.

[12] 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.

[13] Morrison, G.S. "Measuring the Validity and Reliability of forensic likelihood-ratio systems", *Science & Justice,* (51) 3: 2011: 91-98.

[14] http://www.abc.net.au/pm/content/2008/s2451596.htm

[15] Commonwealth Director of Public Prosecutions Report, 2009-2010.