FORENSIC SPEAKER RECOGNITION AT THE BEGINNING OF THE TWENTY-FIRST CENTURY - AN OVERVIEW AND A DEMONSTRATION

Phil Rose

Introduction

How well can someone be recognised by their voice? This is a question that has obvious forensic relevance, and it is being asked more and more often. There is no simple answer, although there is a simple reply: *it depends*. This short paper illustrates some of the main factors upon which forensic speaker recognition (FSR) depends. Its aim is to clarify for interested parties the most important things about the state of FSR at the turn of the century. I will do this by first outlining a typical FSR scenario, and making precise what kind of answer can be given to the question *was the incriminating speech said by the suspect?* I will then give a demonstration with real data of how that answer can be arrived at. Finally, I use this demonstration as a background to a discussion of more general topics of relevance and interest: the evolution of FSR over the last 20 years; the different types of FSR; and the different methods.

The unanswerable question

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

Normally, *this question cannot be answered* by the forensic speaker recognition expert. This is the single most important thing to understand about forensic speaker recognition: everything follows from it. Why? It seems after all such a simple question. The answer is that the identification expert cannot logically say how probable it is that the same speaker is involved unless they are able to take into account not just the voice evidence, but all the evidence in the case. If they are not privy to that evidence, and most of the time they are not, they will be violating *the* basic rule of logical inference if they attempt to say how probable it is that the same speaker is involved, given the speech evidence. This point is so important that it is worth going into in some detail.

Evaluating forensic identification evidence

It is common to represent the important elements in this question in abbreviated form, and we will do this here. Letting "*p*" stand for probability, and "H_{SS}" stand for the **H**ypothesis that the **same speaker** was involved, the probability that the same speaker was involved is represented formally as $p(H_{SS})$. There are several ways that probability can be construed, but

here it is best to understand this formal expression as a number representing degree of certainty. It takes a value between 0 (or 0%), which stands for "certain that it is not the case", and 1 (or 100%), which stands for "certain that it is the case". So, $p(H_{SS}) = 0.95$ (or 95%) can be taken to mean that it is pretty clear to me that the samples were said by the same speaker; $p(H_{SS}) = 0.1\%$ would mean that I am pretty sure that that samples were said by different speakers. And $p(H_{SS}) = 0.5$ (or 50%) means that I cannot say one way or the other.

We also need to incorporate the speech evidence in this formal statement. This will be the inevitably present differences (or similarities, for similarities are just small differences) between the suspect and offender speech samples. Representing the speech evidence by E_{Sp} we include it in the probability statement thus: $p(H_{SS} \mid E_{Sp})$. This is shorthand for "the probability that the offender and suspect speech samples come from the same speaker, given the observed differences between them" (the vertical stroke "|" stands for "given", or "conditional upon").

One final thing needs to be introduced to set up the argument. It looks like a complication, but in fact it makes things easier to understand if instead of using probability we use odds. In their simplest form, odds compare the probability of something happening with the probability of it not happening. So instead of asking *what is the probability that the suspect and offender samples came from the same speaker, given the differences between the speech samples* - $p(H_{SS} \mid E_{Sp})$ - *we ask* ***how much more likely****, given the differences between them, are the samples to have been said by the same person than by different people?*

Expressed as a formula, this becomes: $p(H_{SS} \mid E_{Sp}) / p(H_{DS} \mid E_{Sp})$, where H_{DS} stands, transparently, for the **H**ypothesis that the samples were spoken by **d**ifferent **s**peakers. Odds are thus just a ratio of probabilities. One can convert probabilities into odds and odds into probabilities, but, as just said, it is easier to work with odds.

We thus want to know the answer to $p(H_{SS} \mid E_{Sp}) / p(H_{DS} \mid E_{Sp})$. There is an answer. The solution to this equation has been known for about three hundred years. It is given by Bayes' Theorem. Bayes' Theorem is of paramount importance when one wants to know the probability of a hypothesis given the evidence, and thus it is crucial in forensic identification. It has been styled by a leading scientist in the UK Forensic Science Service as "*...the* fundamental formula of forensic science interpretation" (Evett 1998: 200).

Here, at (1), is the odds form of Bayes' Theorem, suitably subscripted for forensic speaker recognition. It says that the odds in favour of it being the same speaker, given the speech evidence (this is what everyone wants to know and is called the posterior odds and is at the left of the equals sign) can be calculated from two terms: the *prior odds* and the *likelihood ratio.* The prior odds are the odds in favour of the hypothesis before the voice evidence is

$$\frac{p\,(H_{SS} \mid E_{Sp})}{p\,(H_{DS} \mid E_{Sp})} = \frac{p\,(H_{SS})}{p\,(H_{DS})} * \frac{p\,(E_{Sp} \mid H_{SS})}{p\,(E_{Sp} \mid H_{DS})} \qquad (1)$$

Posterior Odds *Prior Odds* *Likelihood Ratio*

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

Suppose a woman is harassed by a sexually explicit phone call from which it is clear that the caller, who can be heard to be an adult male speaker of Australian English, knows her quite well. A sensible estimate of the prior odds is that it could be any one of the number of adult Australian-English speaking males that know her well. If there are, say, about 100 such people, including the suspect, then the prior odds that the suspect is the caller are 99 to 1 *against.* (The *probability* that it is the suspect is 1 **in** 100, or 0.01. That equates to *odds* of (p / [1-p]), or (0.01/[1- 0.01] =) 1 **to** 99.)

The Strength of Evidence

The Likelihood Ratio of Bayes' Theorem is the most important thing in forensic speaker recognition because it is a measure of the *strength of the evidence* in favour of a hypothesis, and it is what the expert should try to estimate. The formula at (1) shows that the Likelihood Ratio too is a ratio of probabilities, but these probabilities are probabilities of *evidence,* not *hypotheses.* Note for example that the numerator is $p(\mathbf{E}_{Sp} \mid \mathbf{H}_{SS})$, not the other way round. The Likelihood Ratio quantifies *how much more likely you are to get the differences between the suspect and offender speech samples assuming they have come from the same speaker than from different speakers.*

Probabilities of evidence

People usually find the idea of the probability of a *hypothesis* relatively easy to understand: e.g. what is the probability that it will rain tomorrow? That the *Raiders* will win the Rugby League Grand Final in '06? That it is the same speaker? They often find the idea of the probability of *evidence* less so. Suppose one part of the speech evidence was that both the suspect and the offender had the same speech defect: saying their "s" sound like a "th", for example (they would then thpeek like thith). Now, proper speech defects, like blood groups or DNA, cannot be controlled at will - the speaker will always say them. So if it was the same speaker in both samples, you would be sure, conditional of course upon there being words with s sounds in them in the samples, to observe that defect, and so p(defect | same speaker) = 1. This is the same as a match between blood found at the crime scene and blood from a suspect. If the blood had come from the suspect then a match would be certain: the probability of getting such a match assuming that the blood had come from the suspect - p(blood group match | same donor) - is one.

Now, what is the probability of the speech defect evidence assuming that the suspect is not the caller, i.e. that the samples came from different speakers? That is the same as the incidence of the defect in the relevant population. If about one in 1000 speakers is known to have the defect, and if someone other than the suspect was involved, you would be likely to get the evidence one time in one thousand, so the probability of the evidence - p(defect | different speakers) - is 1/1000, or 0.001.

If you are more likely to get the speech evidence assuming that the samples came from the same speaker than from different speakers - if $p(\mathrm{E}_{Sp} \mid \mathrm{H}_{SS})$is bigger than $p(\mathrm{E}_{Sp} \mid \mathrm{H}_{DS})$ - that counts as support for the prosecution claim that the samples came from the same speaker. If, on the other hand, you are more likely to get the speech evidence assuming that the samples came from different speakers than from the same speaker - if $p(\mathrm{E}_{Sp} \mid \mathrm{H}_{DS})$ is bigger than $p(\mathrm{E}_{Sp} \mid \mathrm{H}_{SS})$ - that counts as support for the defence claim. If you are just as likely to get the evidence assuming same-speaker as different-speaker provenance - if the ratio of $p(\mathrm{E}_{Sp} \mid \mathrm{H}_{SS})$ to $p(\mathrm{E}_{Sp} \mid \mathrm{H}_{DS})$ is one - the evidence is useless. Thus the magnitude

of the Likelihood Ratio quantifies the strength of the evidence: bigger than unity means support for same-speaker claim; less than unity means support for different-speaker claim; unity (or values close to it) means evidence is useless (or next to useless).

Thus quantified, useless evidence is far from a useless concept. It is often assumed that lack of support for one hypothesis implies support for the alternative. But a Likelihood Ratio of unity means that there is no support for either. Thus, with a Likelihood Ratio of unity it is no good defence claiming that absence of evidence in support of the prosecution claim that the same speaker was involved means that different speakers were involved, and *vice versa.*

Here, finally, is a made-up example of how one can estimate the probability of the hypothesis from prior odds and Likelihood Ratio. Suppose the suspect is one of a group of five males known to be in a house at the time of an incriminating phone intercept, perhaps a bomb threat, from the house. The prior odds (the odds in favour of the hypothesis before the evidence is adduced) are then 4 to 1 *against* them being the owner of the intercepted voice. Suppose further from comparison of known and unknown phone intercepts the evidence is estimated as 100 times more likely if the same speaker is involved (that is, the Likelihood Ratio is 100). The posterior odds on the suspect being the speaker now shift to (prior odds * likelihood ratio = 1/4 * 100/1 =) 25 to 1 in favour. Bayes' Theorem shows how belief in a hypothesis can be updated when new evidence is adduced.

The court must then interpret these odds, or more likely their corresponding probability. If it exceeds some previously determined value - beyond reasonable doubt or the balance of probabilities for example - the defendant is found by the court to have produced the speech samples. In this made-up case $O_{post}(H \mid E)$ = 25:1, which corresponds to a probability of (odds in favour / [odds in favour + odds against] =) 25/26, or 96%. This is clearly beyond the balance of probabilities required in civil cases. Whether it constitutes *beyond reasonable doubt* is up to the court to decide (what a jury construes as beyond reasonable doubt often varies as a function of the perceived severity of the punishment).

The final point in the argument can now be made for why the forensic expert cannot quote the probability of the hypothesis given the evidence. It is this. It is clear from Bayes' Theorem that, *unless the forensic speech recognition expert knows the prior odds, they logically cannot estimate the probability of the hypothesis.* Since the expert is usually not privy to information that informs the prior odds - and in fact there are very good reasons why they should not be (Rose 2002: 64, 74, 273-274) - they cannot logically state the probability of the hypothesis. Since this, in the author's experience, is precisely what is usually expected of the expert by just about everybody involved (instructing solicitors, councel, court and police), this can be a big problem (Boë 2000: 215, Rose 2002: 76-78). It also needs to be acknowledged that this point is sometimes not appreciated even by the practitioners themselves, many of whom still formulate their conclusions in terms of $p(H \mid E)$ (Broeders 1999: 239).

Bayes' theorem, likelihood ratios, and the law

The main textbooks on the evaluation of forensic evidence, e.g. Robertson & Vignaux (1995), or forensic statistics, e.g. Aitken & Stoney (1991), Aitken & Taroni (2004), stress that it is the role of the identification expert to estimate the strength of the evidence by estimating its Likelihood Ratio: the probabilities of the evidence under competing prosecution and defence hypotheses. In a recent review of Aitken & Taroni (2004) it was stated that "The case made for this approach, whether the subject matter is DNA, glass fragments,

clothing fibres or whatever, is overwhelming ..." (Haigh 2005: 40).

It is also possible to find this approach implemented in real case-work, both by experts and the judiciary. For example, in a 1996 appeal court ruling (R v Doheny) concerning expert testimony involving DNA evidence it was stated:

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

This clearly shows that it is the strength of the evidence that is expected of the expert; not a statement of the probability of the hypothesis, given the evidence. In DNA cases, strength of evidence is often expressed with a random occurrence ratio, but this is directly translatable into a Likelihood Ratio. The second paragraph of this quote may well have been motivated by ultimate issue concerns, rather than a knowledge of Bayes' Theorem.

As another example of positive cognisance of the appropriateness of Bayes' Theorem on the part of the judiciary, Hodgson (2002: 109) writes:

> ... it is helpful for courts to be aware of principles of probability theory relevant to the reasoning processes being undertaken. And where evidence is explicitly statistical in character, it really does become necessary to know how such evidence should be integrated with the rest of the evidence in the case. So I believe trial lawyers and judges should have a basic understanding of Bayes' Theorem; and judges should be able to give sound directions to juries in cases, such as some of those involving DNA evidence, where reasoning that in substance gives effect to Bayes' Theorem is required.

In sum, as noted by Broeders at the 14th *Forensic Science Symposium* (2004:173):

> ... recent developments in the interpretation of the evidential value of forensic evidence are now clearly beginning to make themselves felt. Conclusions in the form of a binary yes/no-decision or a qualified statement of the probability of the (prosecution) hypothesis rather than in the form of a statement of the probability of the speech evidence given a set of hypotheses are increasingly criticised for being logically flawed. Instead, conclusion formats that make it possible for results to be expressed in terms of a likelihood ratio are increasingly propagated and becoming more widely used.

Nevertheless, Bayes does not generally seem to be a welcome person in the courtroom. For example, in the very same judgement in Doheny that pointed to the proper statement of the strength of DNA evidence, it was notwithstanding 'strongly endorsed' that :

> To introduce Bayes [sic] Theorem, or any similar method, into a criminal trial plunges the Jury into inappropriate and unnecessary realms of theory and complexity deflecting them from their proper task.

How is this apparent contradiction (the Likelihood Ratio is alright, but Bayes isn't) to be understood? The answer comes from statistics. Although quoting the Likelihood Ratio of the evidence is often styled Bayesian, *the use of a Likelihood Ratio to help in evaluating*

the strength of evidence is not necessarily Bayesian in any special sense (Hand & Yu 2001: 386-7). In formal statistics, the term 'Bayesian' implies, or is associated with, the use of subjective priors (Sprent 1977: 215-6). As just pointed out, legally the priors must not be the concern of the expert witness. Moreover, subjective priors can be anathema in the courtroom, if they ever get that far (Good 2001: 5.5, 6.1, 6.2, 7).

It clear, then, that a crucial distinction needs to be drawn between the forensic use of a Likelihood Ratio to quantify the strength of evidence, which nobody - judiciary or statisticians - can object to, and the additional use of subjective priors. Furthermore, it is clear that the term 'Bayesian' is inappropriate when characterising the FSR approach described in this paper. Since it is the use of a Likelihood Ratio which is crucial forensically, it would be obviously advisable to use a term something like 'Likelihood Ratio-based', rather than 'Bayesian'.

Given the above, it would clearly be difficult to argue now why FSR practitioners should be exempt from the requirement to confine themselves to estimating the strength of evidence and not try to give the probability of the hypothesis. A possible FSR conclusion might thus go something like this. *There are always differences between speech samples, even from the same speaker. In this particular case, I estimate that you would be about 1000 times more likely to get the differences between the offender and suspect speech samples had they come from the same speaker than from different speakers. This gives moderately strong support to the prosecution hypothesis that the suspect said both samples.* To this should probably be added, given everyone's disposition to transpose the conditional, but at the risk of further confusion: *It is important to understand that this does not mean that the suspect is 1000 times more likely to have said both samples. Before you can say how likely it is that the suspect said the incriminating words, you have to take the prior odds into account.*

It is important to realise that there is nothing wrong with the natural question *do you think that the suspect said the incriminating words?* Nothing wrong, that is, as long as the answer - perhaps *yes, the samples sound to me as if they have come from the same speaker* - is treated as evidence for which a LR must be estimated. That is, what would then need to be asked is: what is the probability that you would say that the samples sound as if they had come from the same speaker, given that they had, and given that they had not? In other words, what is the error rate on your gut feeling? You cannot escape the Likelihood Ratio. We will now see how one is estimated from some real speech data.

Demonstration

I will now give an example of how some of the ideas just discussed are put into practice with real data. I have taken recordings from two intercepted land-line telephone conversations, from real case-work, and will compare them using a Likelihood Ratio-based approach. This will involve choosing several acoustic features present in both calls; quantifying the difference between the calls in these features; and estimating how much more probable it is that these differences are same-speaker differences than different-speaker differences. In other words I will estimate the Likelihood Ratio (which will be now abbreviated to LR) for the evidence.

I actually know that both recordings are from the same speaker, answering calls from two different males. The speaker is an adult male with a Broad Australian accent, so the result of the comparison should be a LR bigger than one.

A speaker's voice can change considerably over time (this is called non-contemporaneous variation), and it can also change depending on who they are speaking to, and the formality of the circumstances. These calls were only separated by a few minutes, and were made under similar non-linguistic circumstances: each call is between two males with Broad accents who know each other talking about similar topics. This combination of favourable linguistic and situational comparability should therefore result in a LR quite a bit bigger than one. This comparison can also be taken as an example of a typical test of the method: compare known same-speaker pairs and known different-speaker pairs and see to what extent they are correctly resolved by their LR.

Both recordings contain many tokens of so-called pause particles - when the speaker goes *er*... . Both recordings also contain many tokens of the word *fucken'*, so a typical sentence might be *It's er.. down the fucken' road.* The first vowel in *fucken'*, and the vowel in *er,* have acoustic features, called formant centre frequencies, that relate, among other things, to the size of the speaker's vocal tract and that can be quantified relatively easily. Moreover, there is data available on the distribution of these features in these vowels for Australian males that can be used to estimate the denominator of the LR. We see here, then, two desiderata for choice of features in the forensic comparison of voice samples: there should be many tokens of the feature in both samples under comparison, and there should be some kind of information available to estimate how likely one is to get the feature at random in the relevant population.

Figure 1 shows a spectrogram of the male speaker's *er* pause particle in one of the calls. A spectrogram shows some of the important acoustic details of speech, and it can be a useful tool in court because it shows what was actually measured. The part of the spectrogram corresponding to the *er* is in the right two-thirds, and can be seen to consist of a series of regular-looking vertical striations. The vocal cords vibrate when you say *er*, and these striations correspond to the vibration of the speaker's vocal cords. The left third of the spectrogram is taken up with an inspiratory breath. The bottom axis is time, in seconds. The *er* can be seen to last for about two-thirds of the duration of the spectrogram: about 0.8 of a second. The vertical axis is frequency, in Hertz (Hz), up to 4000 Hz.

A spectrogram shows the amount of acoustic energy present in speech and how it varies over time. The amount of energy is proportional to the darkness of the trace. The most important things in this spectrogram are the three horizontal black bands, labelled F1 through F3, running through the *er.* During speech, the air in the speaker's mouth and throat vibrates at several different frequencies at once. The three bands reflect the frequencies at which the air in the speaker's vocal tract was vibrating when he said this actual *er.* These frequencies are called formants. During this *er,* it can be read off the vertical, frequency, axis that the air was vibrating most strongly at approximately 500 Hz (the first formant frequency), 1500 Hz (F2) and 2500 Hz (F3).

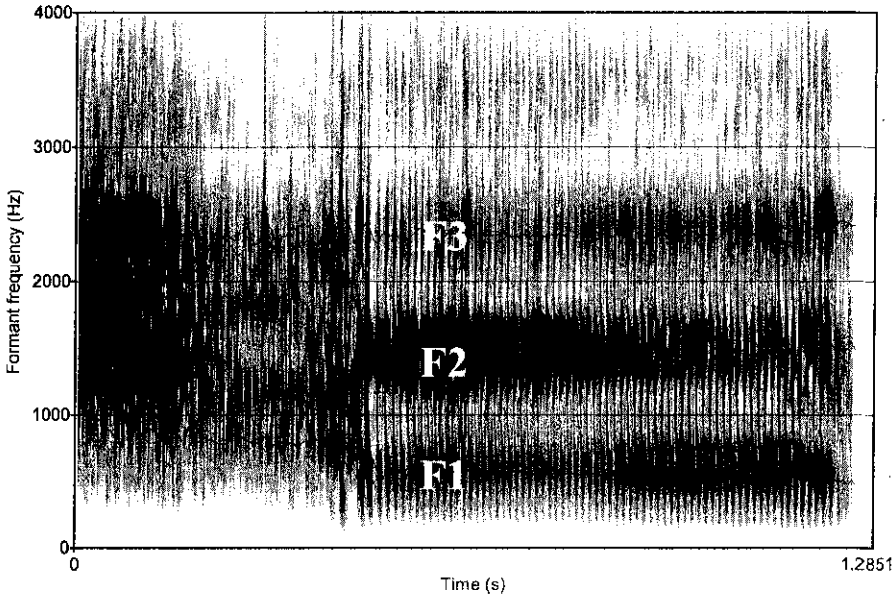

Figure 1. Spectrogram of an Australian male speaker's *er* pause particle from an intercepted telephone conversation. Formants and automatically extracted formant centre-frequencies are shown. F1 = first formant etc.

The formant frequencies are a unique function of the shape of the speaker's mouth and throat, such that if we know the dimensions of the mouth and throat we can predict the frequencies. The dimensions of the speaker's vocal tract depend on the length of their vocal tract, (which is usually correlated with their size), and on the sound they are producing. In figure 1 a computer program has been used to estimate the actual formant frequencies, called the centre-frequencies, and plot them. The automatically estimated formant centre-frequencies can be seen as thin horizontal traces through the middle of the thick formant bands.

The centre-frequencies of the first three formants in *er* are a good example of acoustic features. Their precise values at any point in the utterance, or their average over a given time-span, can be easily, although not unproblematically, obtained from the computer using conventional signal-processing algorithms. Using a fairly steady-state portion of the F-pattern (the ensemble of formants) from sec. 0.578 to sec. 0.859, the average formant centre-frequencies were found to be: 591 Hz (first formant), 1439 Hz (F2), and 2272 Hz (F3). Each one of these will be treated for the moment as a separate feature, such that we could compare two speech samples, for example, with respect to their F2 in *er*, or their F3 in *er* etc.

One has to be careful when interpreting the measurements given by the computer. The F2 and F3 centre frequencies of the *er* vowel are fairly reliably estimated automatically in telephone speech, but the first formant is often a problem. It is well known that the first formant is often seriously affected by the telephone transmission: the speaker's actual F1 frequency is usually shifted higher. This has probably happened in this example, and it is easy to show (although I won't do it here) that the shifted F1 frequency has in addition been estimated too high by the automatic formant extraction (so there are two inaccuracies involved).

This is something the expert simply has to know. Because of this, it is best in this case to simply ignore the first formant as evidence.

This particular token of *er* can be acoustically quantified for forensic purposes, then, by its F-pattern centre-frequencies of 1439 Hz (F2) and 2272 Hz (F3). These values are not invariant, ever. They will vary from *er* token to *er* token. To illustrate this within-speaker variation, figure 2 shows another of the speaker's *er*'s from the same conversation (he is saying *but er* ..). The F-pattern, with its extracted centre-frequencies, can be easily seen. The average centre-frequencies for this token measured over a relatively steady-state portion of the *er* (from sec. 0.279 to sec. 0.465) were found to be: 1410 Hz (F2) and 2362 Hz (F3). It can be seen that these are similar to, but by no means the same as, the values for the previous token.

Suppose, now, we are able to measure some more *er* tokens from this speaker. It would then be possible to estimate the distribution of the F-pattern centre-frequencies of *er* in this speaker's particular conversation. If we were then able to do the same thing for many different conversations, and combine the distributions, we could build up a distribution of the speaker's *er* acoustics over many conversations. This would then allow us to model his *er* F-pattern values as a probability distribution, such that, given a set of questioned F-pattern values, we could *estimate the probability of observing these values assuming that they were said by our hesitant speaker* - i.e. estimate the numerator of the LR: $p(E \mid H_{ss})$.

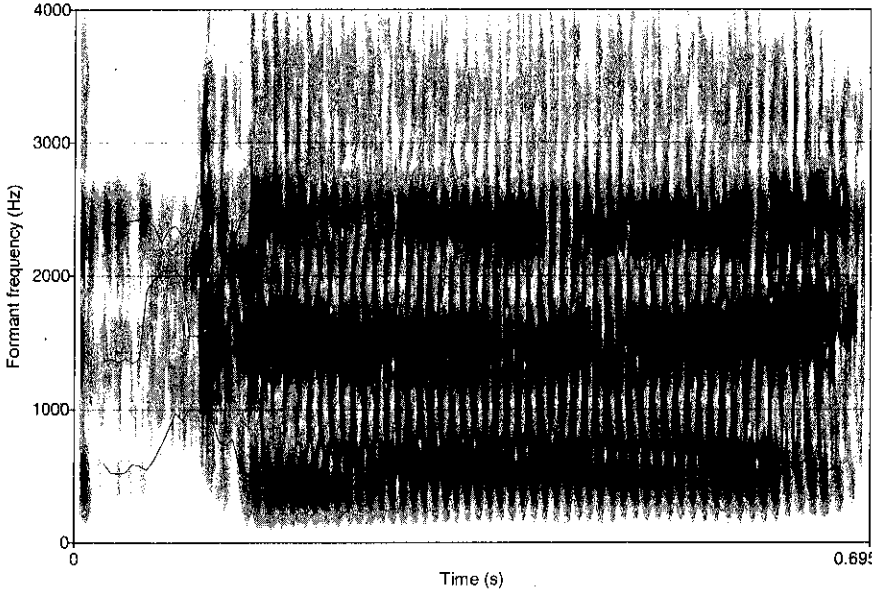

Figure 2. Spectrogram of another *er* pause particle from the same speaker, and the same conversation, as in figure 1.

This is a little too complicated to demonstrate in this paper, and that is why I am giving a simpler demonstration of a LR-based comparison between the *er* F-patterns in two separate conversations from the same speaker. We can pretend that one is the suspect call, and one the questioned.

Fifteen tokens of *er* were measured for F-pattern from the suspect conversation, and also from the questioned conversation. Their means and standard deviations were then calculated and are given in table 1, where it can be seen that the mean values do not differ by very much: 5 Hz for F1; 21 Hz (F2) and 31 Hz (F3). The small difference for F1 is suspicious and, as already explained, is probably because of the telephone transmission and best ignored.

Bayes' theorem shows that similarity is not the only determinant of strength of evidence, however: typicality must also be assessed. We need therefore to know how typical of the relevant population these values are, such that we can estimate how likely we would be to observe them in randomly selected pairs of different speakers. This information is given by a so-called reference, or background, distribution.

Table 1. Means and standard deviations for F-pattern centre-frequencies (Hz) in 15 tokens of *er* in two different intercepted telephone conversations.

Suspect	F1	F2	F3
mean	526	1429	2298
standard deviation	26	30	67
Questioned			
mean	531	1450	2329
standard deviation	20	48	56

Ideal reference distributions in traditional FSR are all but non-existant. This is because not much interest has attached to looking at how a feature distributes in a large number of speakers, and also because it takes a lot of work to amass the data. Fortunately, multi-speaker data are available on Australian English that can serve as reference distributions. One set of data, on the formants of Australian English-speaking males, was painstakingly collected a long time ago, by John Bernard in his 1967 Ph D thesis. We will make use of it in this demonstration.

Figure 3 shows the distribution of the second and third formant in the *er* vowel of a reasonably large number - 57 - of Broad male Australians from the Bernard data set. (The *er* vowel is the vowel that occurs in words like *chirp blurb heard dirt church* etc. I am assuming that this speaker's *er* vowel is in some sense the same as his vowel in the *er* pause particle, although this does not necessarily follow and would have to be justified by additional experimentation.) The formant frequency is shown along the horizontal axis in these distributions. It can be seen that the range of individual F2 values runs from about 1250 Hz to about 1900 Hz, and individual F3 values range from about 2100 Hz to 2800 Hz.

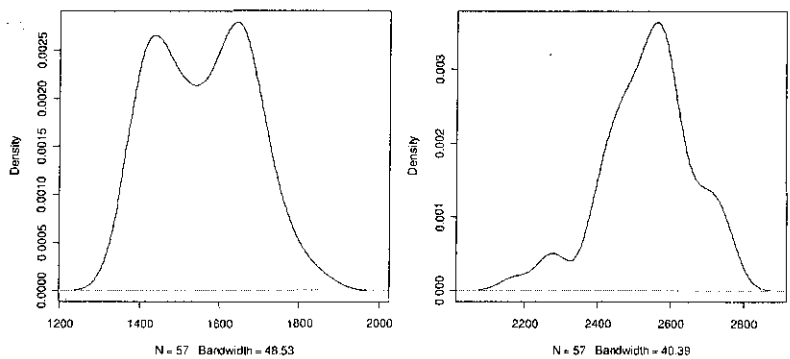

Figure 3. Gaussian kernel probability density functions of the second (left) and third (right) formants in male Broad Australian English *er* vowel. Horizontal axis = formant centre frequency

The distributions in figure 3 can be seen to deviate considerably from normal, or so-called Gaussian distributions: F2 is clearly bimodal (has two peaks), and F3 is slightly positively skewed. This deviation introduces complications. If the background distributions do not deviate excessively from normality, they can be modelled by normal distributions without much loss of accuracy in the LR estimation. Here, the deviation from normality means a LR estimate assuming normality would not be accurate. Consequently, the evaluation of the LR will involve a more complicated formula which takes the non-normal distribution of the background features into account.

A further complication is that the two features of second and third formant frequencies in *er* are not independent. We can observe this empirically from the moderate degree of positive correlation (0.38) between F2 and F3 in the reference population of General Australian Male *er* vowels. We also know it analytically, from phonetic theory. Given this particular *er* vowel, it is known from the received Acoustic Theory of Speech Production that F2 can be predicted from F3 or *vice versa*. It is assumed that an *er* vowel is produced with a vocal tract of uniform cross-sectional area, and the formant frequencies of such a tract will be a function of the length of the tract. (The frequency F*n* of a formant *n* from a tract with this configuration is given by $Fn = (2n\text{-}1)^* (C/4l)$, where *l* is the length of the vocal tract in centimetres and C is the speed of sound in cms./sec.)

According to theory, therefore, an observed F2 of 1429 Hz for an *er* vowel, as in the first conversation, would be associated with a vocal tract 18.37 cms. long, and a vocal tract this long would produce an *er* with an F3 of 2382 Hz. This value differs from the observed F3 by just 84 Hz, so the prediction is quite good. Going the other way, an observed F3 of 2298 Hz would predict an F2 of 1379 Hz, another small difference of 50 Hz.

If the F2 and F3 frequencies had been independent, Bayes' Theorem says that their combined evidential strength could have been found by simply taking the product of their individual LRs. Since the formula for univariate LRs is less complicated, this would have been nice. But the fact that F2 and F3 in *er* are positively correlated means that only *some* extra information is being provided by taking both frequencies into account in the comparison, and it is certainly not legitimate to present their combined evidence as the simple product of their two separate LRs. For this reason a more complicated LR formula has to be used: one that can take the correlation into account. We use here a so-called *two-level bivariate kernel density LR* formula. It is a version of a formula for the multivariate-normal LR derived in Aitken & Lucy (2004), where it simplifies to the expression at (2). *Two-level* means that the formula is taking two levels of variance into account: the variance in a feature *between-speakers*, and the variance *within* a given speaker on the same occasion. Since the conversations were close in time, a third level of variance: that within a given speaker on *different occasions*, has not needed to be included. *Bivariate* means that only two variables, or features, are being compared. *Kernel density* refers to a method of modelling the reference distribution assuming non-normality, the need for which was explained above.

Although it looks complicated, and it is, it is important to remember that the Likelihood Ratio formula at (2) is still just comparing the similarity of the two mean F2 and F3 measurements in the suspect's and offender's speech with their typicality against an appropriate reference population.

$$\mathrm{LR} = \frac{\left|2\pi\left[(n_1+n_2)U^{-1}+C^{-1}\right]^{-1}\right|^{-1/2} exp\left\{-\frac{1}{2}(H_2+H_3)\right\}}{\left|2\pi C\right|^{-1/2}\left|2\pi\left(n_1U^{-1}+C^{-1}\right)^{-1}\right|^{-1/2}\left|2\pi\left(n_2U^{-1}+C^{-1}\right)^{-1}\right|^{-1/2} exp\left\{-\frac{1}{2}(H_4+H_5)\right\}} \quad (2)$$

U = Within-group covariance matrix, C = between-group covariance matrix,
n_1, n_2 = number of replicates in offender and suspect samples
$H_2 = (y^* - \mu)^T ((U/(n_1+n_2)) + C)^{-1} (y^* - \mu)$,
$H_3 = (\bar{y}_1 - \bar{y}_2)^T (D_1+D_2)^{-1} (\bar{y}_1 - \bar{y}_2)$,
$H_4 = (\mu - \mu^*)^T [(D_1+C)^{-1} + (D_2+C)^{-1}] (\mu - \mu^*)$,
$H_5 = (\bar{y}_1 + \bar{y}_2)^T (D_1+D_2+2C)^{-1} (\bar{y}_1 - \bar{y}_2)$,
$y^* = (n_1\bar{y}_1 + n_2\bar{y}_2) / (n_1 + n_2)$,
$\mu^* = \{(D_1+C)^{-1} + (D_2+C)^{-1}\}^{-1} [(D_1+C)^{-1} \bar{y}_1 + (D_2+C)^{-1} \bar{y}_2]$,
$D_1 = (1/n_1)\ U$, $D_2 = (1/n_2)\ U$

The LR formula at (2) estimates how much more likely you would be to get the difference between the means of the suspect and offender samples assuming that they had come from the same speaker than assuming they had come from different speakers taken at random from the reference population.

To make it easier to understand what is involved, the LR-based comparison of the two suspect and offender means against the reference population is shown graphically. We start with the distributions of F2 and F3 in the suspect and offender samples: these are shown in figure 4. In figure 4, two features are being shown combined (this is what is meant by *joint* distributions), so the figure is three-dimensional. The F2 values are shown along the bottom; the F3 values are shown increasing up the left-hand side, and the vertical dimension shows their probability density (this is not the same as probability, but for the purposes of the demonstration it can be understood as such). The distributions have been modelled normally. This means they can each be constructed with just information on the mean and standard deviation of both features, and number of items in the sample, together with the amount of correlation between both features.

The two distributions can be seen to intersect (the mean and standard deviation values in table 1 show they are, after all, very similar): the bottom part of the offender's distribution can be seen sticking out of the suspect's distribution in the bottom left-hand corner, and one can also see a bit of the offender's distribution on the other side of the suspect. Interestingly, the two distributions show different correlation between F2 and F3. The offender's distribution shows F2 and F3 positively correlated: it is oriented bottom-left-to-top-right, showing F3 increasing with F2. The suspect's distribution on the other hand shows F2 and F3 slightly negatively correlated: F3 decreases with increasing F2.

Protruding from the apex of both distributions is a line with a ball atop. This shows the location of the samples' mean F2 and F3 values. One can imagine the lines being continued downwards inside the distributions to meet the F2 / F3 plane at the intersection of 1429 Hz (F2) and 2298 Hz (F3) for the suspect, and slightly higher at the intersection of 1450 Hz (F2) and 2329 Hz (F3) for the offender.

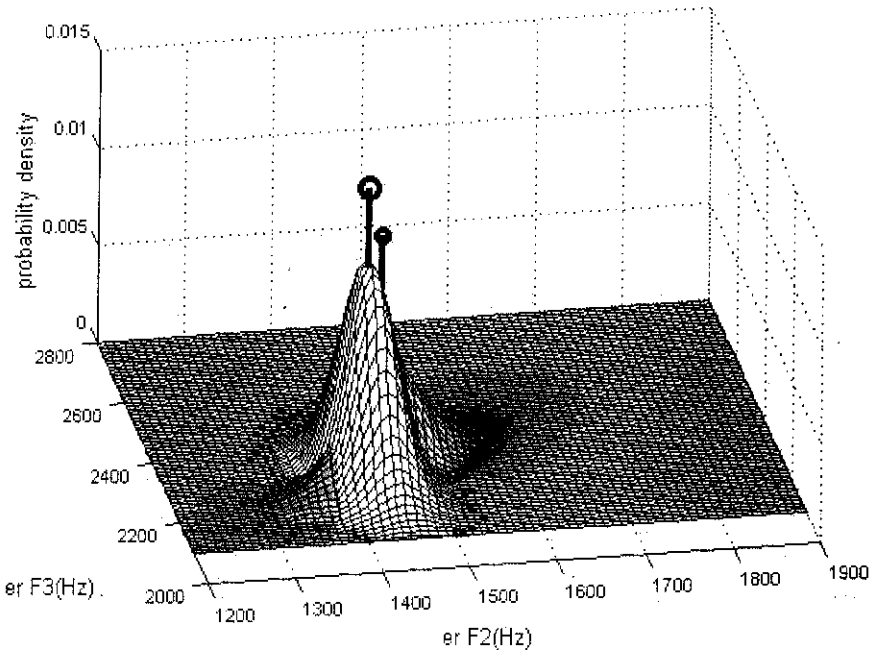

Figure 4. Joint bivariate normal probability density distributions for F2 and F3 in suspect and offender's *er* vowels. Vertical lines with balls on top show location of means of suspect (large ball on top) and offender (small ball).

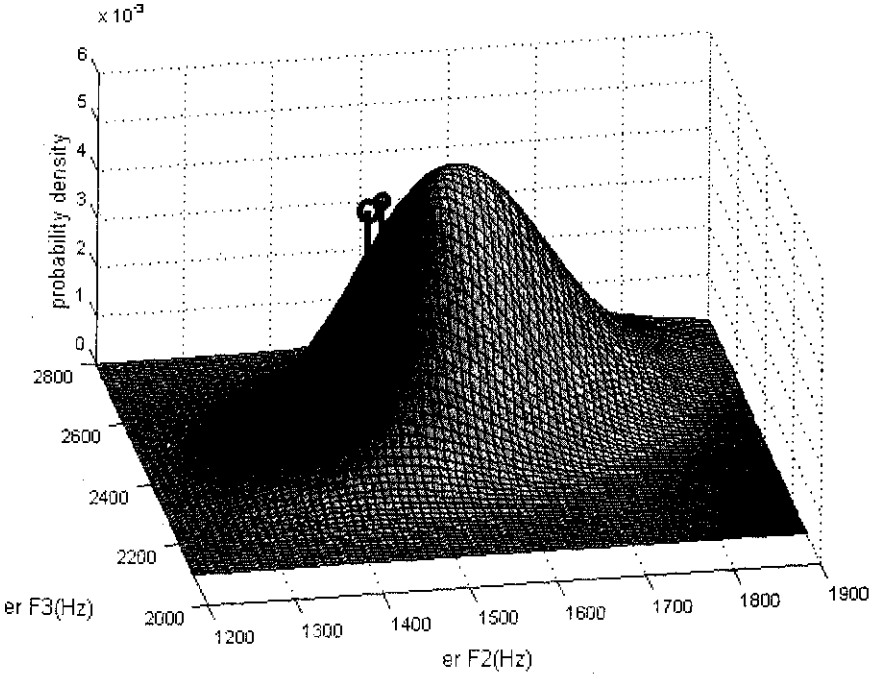

Figure 5. Joint bivariate normal probability density distribution for F2 and F3 in Broad Australian *er* vowel from 57 speakers. Vertical lines show location of means from suspect (large ball on top) and offender (small ball).

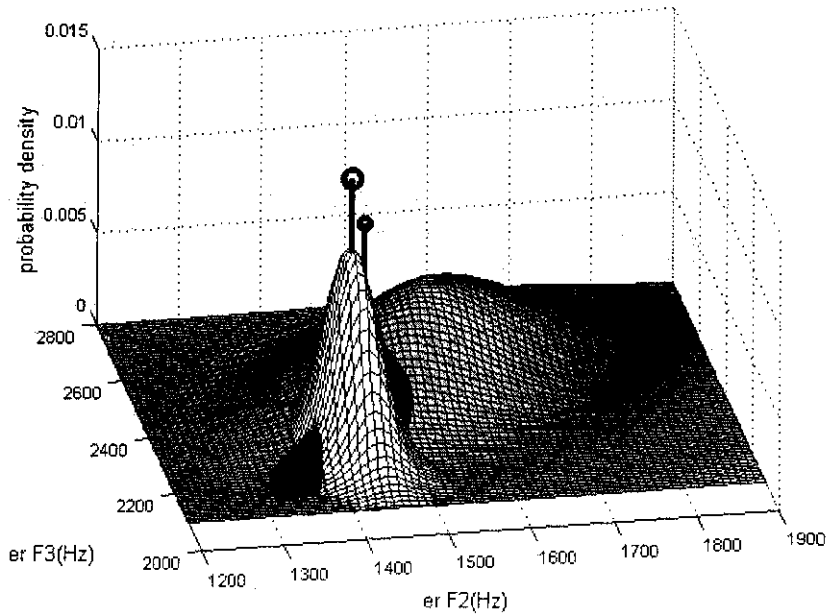

Figure 6. Joint bivariate normal probability density distributions for F2 and F3 in reference population, and in suspect and offender samples. Vertical lines show location of means from suspect (large ball on top) and offender (small ball).

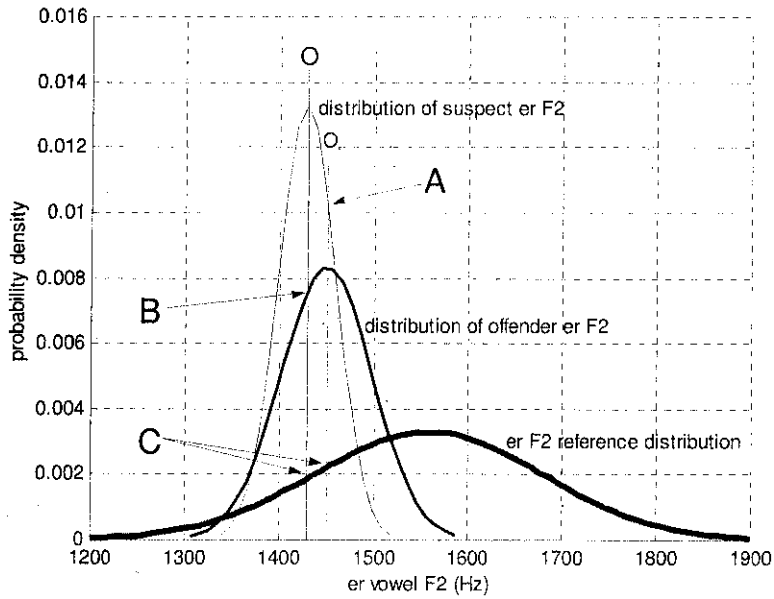

Figure 7. Univariate Likelihood Ratio-based comparison of suspect and offender *er* vowel F2. S, O = location of suspect and offender *er* F2 means. A = probability density of observing offender mean (1450 Hz) assuming it came from suspect; B = probability density of observing suspect mean (1429) assuming it came from offender; C = probability densities of observing suspect and offender means, assuming they have come from the reference population.

Now we examine the distribution of F2 and F3 in the reference population - 57 Broad Australian males. This is shown in figure 5. For simplicity, the distribution has been modelled normally, but it should be remembered that the distribution of F2 and F3 in the population was shown above to deviate from normality, and in reality it would be a bit more uneven and bumpy. F2 can be seen to range from 1250 Hz to 1850 Hz, and F3 from 2100 Hz to 2800 Hz.

The locations of the suspect and offender mean values are shown, as in figure 4, by vertical lines with balls on top. It can be seen that they are both fairly atypical of the distribution, lying towards the bottom left-hand corner. Given the bimodal distribution of F2 in *er,* the values might not be quite so atypical as implied by the normal distribution.

Figure 6 combines the distributions in figures 4 and 5, showing the distributions of suspect and offender samples against the distribution of the reference population. The reference population distribution appears flatter than the suspect and offender distributions because all the distributions are probability distributions and must have unit volume. Because the reference population has a wider spread in two dimensions, its peak cannot be so high.

It is still difficult to give the idea of a LR estimation in two dimensions, so we now reduce the comparison to just F2. Imagine you are looking just at the F2 side of figure 6. Figure 7 shows what you would see. The horizontal axis of figure 7 shows the centre frequency of F2 in *er,* and runs from 1200 Hz to 1900 Hz. The vertical axis shows the probability density. This is not the same as the probability, but for the purposes of LR estimation, and the purpose of demonstrating it in this figure, the probability density can be taken to reflect the probability of getting a particular *er* F2 value, given its probability distribution.

Figure 7 contains three distributions. The flat one with the thick line is the distribution of the reference population, which shows us how the feature (the F2 of the *er* vowel) distributes in the selection of 57 Broad male Australian English speakers. It is centred around about 1550 Hz. This enables the estimate of the typicality of the observations. The other two, more peaked, distributions are of the suspect and offender *er* F2. The offender *er* F2 has a mean value of 1450 Hz. Its location is shown by the vertical line surmounted by the small "O". The suspect *er* F2 has a mean value of 1429 Hz, shown by the line and larger "O".

Point A in the figure shows the probability (density) of observing the offender's mean value of 1450 Hz assuming it came from the suspect: (note: probability of *evidence* given hypothesis!): it is about 0.0105. Point B shows the probability (density) of getting the suspect's mean value of 1429 Hz, assuming it came from the offender: it is about 0.0075. Points C show the probability (density) of getting the suspect's and offender's values assuming they have been taken at random from the reference population: they are both near to 0.002. It can be seen that the probability (density) of getting the values assuming they have come from the population is smaller than the probabilities that the suspect and offender samples have come from the same speaker. It is about four and a half times smaller, in fact: $[(0.0105 + 0.0075)/2] / 0.002 = 4.5$. That is the LR for this evidence: 4.5. It is bigger than one: the probability of the evidence assuming same-speaker provenance is greater than the probability of the evidence assuming different-speaker provenance. This will therefore support the prosecution hypothesis. But only a little, because 4.5 is not a large LR.

Now imagine that the same was done for F3, but both LR values combined in such a way as to take their correlation into account, using formula (2). The result is a LR of 32.8. This means that, when evaluated against the reference distribution, the difference of 21 Hz between the suspect's and offender's mean F2 centre frequencies in *er* and the difference of 31 Hz between their F3 values would be about 33 times more likely were the samples from the same speaker than from different speakers. Since we in fact know that they are from the same speaker this is an encouraging result.

Given the way that the features in voices vary, basing a forensic comparison on just one or two features is not tenable. In the comparison of voice samples from the same speaker, as here, there are bound to be examples where the difference between the features is evaluated as more likely given that they have come from different speakers rather than the same, thus contributing a LR less than one and in support of the defence. The same applies *mutatis mutandis* for comparison between different-speaker samples: you will almost certainly get LRs bigger than one, and thus more typical of same-speaker data, for some features. This is simply the way voices are. It is therefore essential in the forensic comparison of voice samples to compare as many different features as possible, and to try to choose those which are likely to be maximally independent of each other. This will have the effect of increasing the probability that the combined LR will be forced either well above or well below unity, thus ultimately contributing to useful strengths of evidence. This is illustrated by now taking into account features from the formants in the word *fucken'*.

Fucken' tends to occur a lot in male forensic voice samples. That is useful, since the magnitude of the LR is also a function of the number of items in the sample, and the more *fucken*'s there are, the greater the probability that the LR will achieve a useful value. In the two conversations from the same speaker, there was, fortunately, a lot of *fucken*'s. I selected 35 tokens in each conversation, and measured the F2 and F3 in the short *ah* vowel in the first syllable of each *fucken'*

Figure 8 shows a spectrogram of one of the *fuckens*, in the utterance the *fucken' work*. The second and third formants in the *ah* vowel in *fuck* are marked. It can be seen that the formants are not static over time, but show considerable movement. The second formant, for example, rises in frequency from about 1200 Hz to 1500 Hz, and at the same time the third formant drops in frequency from about 2300 Hz to 1800 Hz. The abrupt changes in formant frequency reflect the rapid changes in the speaker's vocal tract shape as it moved from the *f* sound at the beginning of *fuck* through its *ah* vowel towards the *k* sound at the end of the syllable. In order to most accurately sample the frequencies of the vowel target, the formant frequencies were measured in the middle of the duration of the *ah* vowel (this middle sampling point can be determined from other features in the spectrogram).

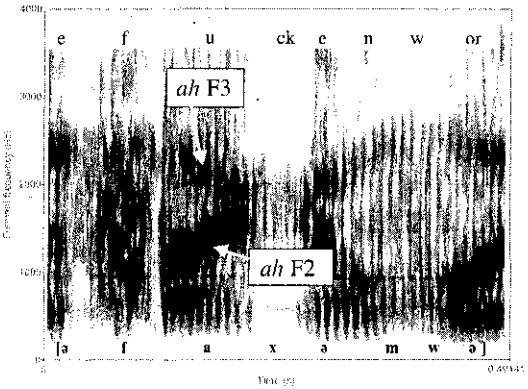

Figure 8. Spectrogram of *(th)e fucken' wor(k)*, showing second and third formants in the *ah* vowel in *fucken'*. The approximate location of the words in the acoustics is shown in spelling along the top. The bottom shows a broad phonetic transcription of the utterance in the International Phonetic Alphabet.

When evaluated against F2 and F3 values for the *ah* vowel in 57 Broad male speakers, using formula (2) in the same way as with the *er* vowel, a LR of 70.4 is obtained. The difference between the suspect's F2 and F3 means in his *ah* vowel, and the offender's F2 and F3 means in his *ah* vowel, is about 70 times more likely were the suspect and offender to be the same person. Once again, this LR is bigger than one, and so is consistent with the known fact that the same speaker is involved.

Now the results for the two vowels have to be combined. F2 and F3 in *ah* and F2 and F3 in *er* are also going to be correlated, not only because they come from the same speaker within a sample, but also because they encode values along the same linguistic feature: both *ah* and *er* are phonetically central vowels, and phonetic theory says that F2 encodes centrality, so one would expect F2 in *ah* and *er* to be correlated.

This means that the combined LR from the *er* vowel and the *ah* vowel acoustics is not going to be simply the product of their two individual LRs of (32.8 x 70.4 =) 2,309, but a value somewhat lower.

How much lower? It is not clear at present from forensic statistics how best to estimate the amount by which this particular LR should be reduced, given an assumed degree of positive correlation between the acoustics of the two vowels. Ideally, one would like to process all four variables - F2 in *er* and *ah*, and F3 in *er* and *ah* - together, as was done for the F2 and F3 variable in the bivariate analysis above. However, whereas F2 and F3 in *er* can be treated as *bona-fide* multivariate data - one can characterise each *er* vowel in the data in terms of these two variables - and F2 and F3 in *ah* are similarly multivariate data, all four variables together are not multivariate. That is, one cannot think of an entity, at this level of analysis, such that it could be characterised by these four variables together, in the same way that an *er* vowel token can be characterised by its F2 and F3. There are, for a start, different numbers of *er* and *ah* vowels involved in the comparisons - 15 *er* vowels and 35 - *ah* vowels, so it is not possible to pair them up. Such a multivariate approach would only become possible if we had more data allowing us to estimate the correlation between mean values across many different conversations of the same speaker.

One way around this is to run experiments to see what the reduction in LR would be, *were* the four variables truly multivariate. That is, one could take the same *er* and *ah* formant data, and calculate LRs for their comparison assuming they were truly multivariate. One could then compare those values with LRs based on taking the product of two bivariate LRs, under the assumption they were independent, and note the amount of reduction in LR, if any. The observed LR value in the present case could then be scaled down according to the results of the control experiment. This experiment was in fact run on two sets of data largely comparable with the data under comparison (one set was from the *er* and *ah* vowels of 11 male speakers of General Australian; one set from the *er* and *ah* vowels of 53 male speakers of Broad Australian). It was found that for one set of data an "independence" LR value of 2309 corresponded to a multivariate LR of 522, and for the other set a value of 1276 was obtained. Note that, as expected, a reduction in LR was observed, but that the result was still larger than the largest individual LR of 70. Thus a sensible estimate of the true LR for our comparison is that it lies somewhere between about 500 and 1200. The conservative estimate is taken, to favour the defence. Which means that we estimate that one would be at least about 500 times more likely to get the difference between suspect and offender bivariate *er* and *ah* samples were they to have come from the same rather than

different speakers. Using the verbal equivalents for LRs proposed for the UK Forensic Science Service (Champod & Evett 2000: 240), this would then constitute 'moderately strong' support for the prosecution hypothesis that offender and suspect samples had come from the same speaker.

If this were a real case, one would then go on to estimate LRs for whatever other useful features there are in the calls. Some of these LRs would be smaller than one and have the effect of reducing the combined LR. For example, if we estimated a LR of 0.5 for a given feature (which means (1/0.5 =) twice as likely to get the difference between the features in suspect and offender samples if a different speaker were involved), this would reduce the combined LR by a half. However, if a sufficiently large number of features can be compared, say, another 20 features, this will have the effect of making it probable that the LR will be forced further and further away from unity, and a clear picture will emerge of the strength of the evidence.

With its use of spectrograms, speech acoustics, complicated formulae and statistics, this analysis may appear scientific, and so it is. There can be no doubt that this is the correct approach to estimating the strength of evidence. But it is also still very crude, and it needs to be stated in what way. The list is long. The LR formula, despite its complexity, does not adequately reflect the complexity of speech and makes very many simplifying assumptions, for example, that all variances are equal. They are not: different speakers can show different amounts of variance, and the same speaker may vary differently on different occasions. Then there is the problem of incorporating more than two levels of variance. When one attempts to combine more than two levels of variance and correlation between features and unequal variances and non-normal distributions - all of which should be assumed in normal forensic comparison - the problem of deriving an appropriate LR formula becomes non-trivial. Forensic statisticians are still working on it, but not yet with data as complicated as speech. (For an example of an LR-based approach which attempts to take correlation into account in FSR, see Rose et al. 2004). The upshot of this is that LR estimates still cannot pretend to very great accuracy.

Given these limitations, it is very important to be able to show that the approach actually works with the formulae at hand. This is at once both the hallmark of a proper scientific theory: that it is capable of being tested; and an important desideratum in the wake of the *Daubert* ruling that scientific evidence needs to have been tested as one criterion for admissibility (Daubert 1993). The extent to which the approach has been tested will be addressed below, after a brief discussion on different types of forensic speaker recognition.

Types of forensic speaker recognition

There are several different types of forensic speaker recognition, and it is important to distinguish them because they are associated with different strengths of evidence.

The first main distinction is between *Technical* and *Naïve* Forensic Speaker Recognition (Nolan 1983: 7). The latter refers to peoples' unreflected ability to recognise voices, and can be found for example in voice line-ups, or when the police claim to recognise the voice of an offender over the phone as that of the suspect.

Technical FSR on the other hand refers to the use of theories and axioms from well-established disciplines like Linguistics, Phonetics, Acoustics, Signal Processing and Statistics. It is also called Expert FSR (Broeders 2004). The demonstration above was an example of Technical, or Expert, FSR.

Two types of Naïve FSR must be carefully distinguished: *familiar* and *unfamiliar*. Familiar FSR is when the person making the recognition is familiar with the voice in question: it belongs to their husband, a good friend, a media personality, for example. People are usually very good at recognising familiar voices (Rose and Duncan 1995), and the associated LR for the evidence must be non-negligible. But familiar performance is not without error: how many times have you thought you were speaking to someone you knew on the phone only to realise you had made a mistake because that is who you expected to hear? Because people differ in their ability to recognise voices, and because recognition depends clearly on a host of other factors - for example how much speech the listener heard, or how distinctive the voice is (voices differ considerably in their distinctiveness), the accurate estimation of the strength of evidence in such cases is extremely difficult (Rose 2002: 97-107). Unfamiliar Naïve recognition is so prone to error that it is of dubious use: its associated LR will not be very different from unity. There are many injunctions in the relevant literature that it should not be used forensically (Rose 2002: 99-100).

There are two different approaches to Technical FSR: *Automatic* and *Traditional.* The difference is addressed in detail in Rose (2003, 2005). The demonstration above with *er* and *ah* was Traditional. It used features (like formants) that can be clearly related to speech production and perception, and can be explained using concepts from linguistics. Traditional features can be either *acoustic*, as above, or *auditory.* Auditory features are those that can be extracted by an expert trained in phonetic transcription and linguistic analysis. They do not necessarily have to do with sound, and can refer to non-sound features like word or sentence structure. It is the consensus that a Traditional approach must involve both auditory and acoustic analysis.

Automatic methods were originally motivated, of course, by a desire to avoid so-called subjectivity in forensic analysis. However, although human interaction is kept to a minimum in order to maximise so-called objectivity (there is no such thing as objectivity in science), there is in fact a lot of human judgement involved. The term *Automatic* is thus a little misleading.

It is important to distinguish Automatic from Traditional approaches because they are associated with different strengths of evidence. Other things being equal, Automatic approaches are far stronger: they will, on average, yield LRs deviating much more from unity than Traditional approaches (Rose 2006). They can also take into account the whole of the speech available, rather than work with a few features extracted from the samples, as with a Traditional approach. This is a big advantage.

But there is no such thing as a free lunch in FSR. What the Automatic approaches gain in *strength* they lose in *interpretability*: they use features from advanced signal processing which are very difficult, perhaps impossible, to explain to a lay audience. Perhaps the main drawback of Automatic approaches is that they are also extremely channel-sensitive: it would be very difficult, for example, to compare a police interview recording of the suspect made directly onto a cassette tape with an intercept of the offender's voice from a mobile phone. Questioned, suspect and reference data have to been recorded under exactly the same conditions for the comparison to work properly, and often it is not possible to even find out what the conditions of the recordings were - for example what kind of data compression, if any, was used. And automatic approaches, by definition, do not take into account linguistic features which can be of evidentiary value. All this is probably behind Broeders' rather negative appraisal of Automatic methods at the 2004 14th *International Forensic Science Symposium:*

> In spite of the regular appearances of high-tech speaker identification equipment in contemporary fiction and the film industry, forensic speaker identification at the beginning of the 21st century is an extremely challenging field, in which the promise held by technological advance remains largely unfulfilled. (p.174).

However, both approaches have their strengths and weaknesses, and are complementary. In fact, if it is conceded that the aim of FSR is to estimate the strength of evidence, it is clear than both approaches are failing to take some evidence into account, and so should be combined (Rose 2006). An example combining aspects of both Automatic and Traditional approaches can be found in Rose et al. (2002), and in fact Automatic approaches are beginning to find an improvement in performance if they include some Traditional features. A complete and proper integration of both Traditional and Automatic approaches is clearly the way to go.

Historical background

The current state of FSR - what is known about how to discriminate between same- and different-speaker samples, and how that knowledge is implemented - is due to the intersection of several historical factors, the two most important of which are the resurrection of Bayes' Theorem and the *Daubert* ruling.

The beginning of FSR can be reasonably dated from its first institutional use, by the Bundeskriminalamt, and with Traditional approaches, in 1980. It was also used, Traditionally, in the 80's and 90's by individual practitioners in the UK and Australia for example. However, the absence of a coherent theory for expressing the strength of evidence meant that it was never clear what to actually do with the observations. There were always differences between forensic speech samples, and the FSR expert was expected to say how probable it was that the samples had come from the same speaker, given the differences between the samples: i.e. a $p(H \mid E)$ statement. During this period, there was little actual testing of FSR approaches, because there was no theory to guide the questions. There had, to be sure, been plenty of research into speaker recognition, but it had not been appropriate for the forensic context.

The advent of DNA profiling in 1985 and its subsequent statistical evaluation, together with some spectacular miscarriages of justice due to incorrect statistical reasoning, has, over the last twenty or so years, focussed considerable attention on the proper, rationalist evaluation of forensic evidence (Dawid 2005: 6). In particular, attention has been drawn to the central role of the LR of Bayes' Theorem in quantifying the strength of forensic evidence (Robertson and Vignaux 1995).

An awareness of the LR and its role in Bayes' Theorem can be considered a true watershed in FSR. That Bayes' Theorem could be applied to FSR was actually first mentioned in 1984. No-one apparently took much notice, because it took some 14 years more before it percolated through to the speech community. The next published reference to it was in Rose (1997), but by that time researchers in automatic speaker identification in Europe had already started actual testing, and the first real demonstration of the approach in forensic speaker recognition research, by Meuwly et al., occurred in 1998. Same-speaker and different-speaker pairs were compared in ways similar to the *er/fucken'* example above, but with Automatic features, to see to what extent same-speaker pairs were resolved, as Bayes' Theorem says they should, with a LR greater than 1, and different-speaker pairs resolved with LRs smaller than one. The result was successful, and Bayes took off in automatic FSR. Meanwhile, in Australia, Kinoshita in her 2001 PhD thesis was the first to successfully demonstrate the method with Traditional features.

The first text-book on FSR for a general audience advocating a LR-based approach appeared in 2002: the author's *Forensic Speaker Identification* in the Taylor and Francis

Forensic Science series. This was followed in 2003 by his monograph on *The Technical Comparison of Forensic Voice Samples.* This was in the legal reference series *Expert Evidence,* and was written for legal professionals.

Subsequently, research has focussed on improving the discrimination of the method, and addressing the theoretical problems of estimating LRs from speech (e.g. Gonzalez-Rodriguez et al. 2006). The most recent developments in Automatic FSR have been the blind testing conducted by the Netherlands FSI using real forensic material, the positive results from which are reported in Leeuwen and Bouten (2004). The most recent results with Traditional methods, again in Australian English, can be found in Alderman (2004a,b, 2006). An up-to-date summary of types of FSR evidence and the results of testing the LR-based approach can be found in Rose (2006).

It would be wrong to deduce from the foregoing that the Likelihood Ratio-based approach to FSR is everywhere accepted and well-established. The degree to which the approach is used, or even understood (for it is difficult to understand), differs from country to country: see Rose (2002: 67-78, 2006) for details. There is no doubt, however, that given the increasing interest in the correct evaluation of forensic identification evidence in general - see the quotes above - one ignores the Likelihood Ratio in Forensic Speaker Recognition at one's own peril.

Summary

This paper has discussed some important aspects of forensic speaker recognition. It has emphasised that the task of a forensic speaker recognition expert is, after first quantifying the differences or similarities between the samples they are comparing, to estimate how much more likely this evidence is, assuming the samples have come from the same speaker than assuming they have not. The paper has, using Bayes' Theorem, explained why this is so, and shown how it is possible to do it, with a real example. It has also taken care to point out the shortcomings in the approach. There are shortcomings in the statistical modelling, which, although already highly sophisticated, is still not quite up to the complexities of speech, and there are also shortcomings in the availability of true reference populations.

It should be clear from the paper that, properly done, FSR is a very complicated matter involving expert knowledge of, at least, linguistics, acoustics, statistics and signal-processing. It is not, as quite commonly supposed, a touchy-feely exercise whereby some individual gifted in recognising people by their voice listens to the recordings and makes their decision. It is also a painstaking, time-consuming, and, given the content of male telephone conversations in general, not very exciting undertaking. (The measurements for the *er/fucken* analysis demonstrated above took about ten hours. It took less than a second to press the key for the Likelihood Ratios, but a very long time to write the programs that derived them. The experiments to estimate the amount of reduction in LR for assumed correlated data took about three days.)

Most important of all, however, the FSR expert needs to know how to interpret their findings forensically. This paper has shown how the Likelihood Ratio of Bayes' Theorem is now considered the proper construct for these findings - indeed, estimating the probabilities of the evidence under both prosecution and defence hypotheses must structure the whole forensic speaker recognition approach, as it should.

REFERENCES

Aitken CGG: *Statistics and the Evaluation of Evidence for Forensic Scientists.* Wiley, Chichester, 1995.

Aitken CGG, Lucy D: Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53/4: 109-122, 2004.

Aitken CGG, Stoney DA: *The Use of Statistics in Forensic Science.* Ellis Horwood, Chichester, 1991.

Aitken CGG, Taroni F: *Statistics and the Evaluation of Evidence for Forensic Scientists.* [2nd ed. of Aitken 1995]. Wiley, Chichester, 2004.

Alderman T: The Use of Australian-English Vowel Formant Data Sets in Forensic Speaker Identification. In Cassidy S. editor. *Proc. 10th Australian Intl. Conf. on Speech Science and Technology* (PANZE workshop): 177-182, 2004a.

Alderman T: The Bernard Data Set as a Reference Distribution for Bayesian Likelihood-Ratio-based Forensic Speaker Identification using Formants. In Cassidy S. editor. *Proc. 10th Australian Intl. Conf. on Speech Science and Technology:* 510-515, 2004b.

Alderman T: *Forensic Speaker Identification: A Likelihood Ratio-based Approach Using Vowel Formants.* LINCOM Studies in Phonetics 01, Lincom Europa, Munich. 2005.

Bernard JRL: *Some measurements of some sounds of Australian English.* Unpublished Ph.D. Thesis, University of Sydney, 1967.

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

Broeders APA: Some observations on the use of probability scales in forensic identification. *Forensic Linguistics:* 6/2: 228-41, 1999.

Broeders APA: Forensic Speech and Audio Analysis Forensic Linguistics - A Review: 2001 to 2004. 14th International Forensic Science Symposium, Lyon. http://www.interpol.int/Public/Forensic/IFSS/meeting14/ReviewPapers.pdf 2004.

Champod C, Evett I: Commentary on Broeders (1999). *Forensic Linguistics,* 7/2:238-43, 2000

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

Evett IW: Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice,* 38/3: 198-202, 1998.

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* Special Issue, 20(2-3): 331-355, 2006.

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

Haigh J: Review of Statistics and the Evaluation of Evidence for Forensic Scientists. *Significance,* March: 40, 2005.

Hand DJ, Yu Keming: Idiot's Bayes - Not So Stupid After All? *International Statistical Review,* 69/3: 385-398, 2001.

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

Kinoshita Y: *Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants.* Unpublished Ph.D. Thesis, the Australian National University, 2001.

Leeuwen DA, Bouten JS: Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation. In: In Ortega-García J et al. editors: 81-82, 2004.

Lindley DV: Probability. In Aitken & Stoney, editors: 27-50, 1991.

Meuwly D, El-Maliki M, Drygajlo A: Forensic Speaker Recognition using Gaussian Mixture Models and a Bayesian Framework. COST-250 Workshop, Ankara, 1998.

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

Ortega-García J, González-Rodríguez J, Bimbot F, Bonastre J-F, Campbell J, Magrin-Chagnolleau I, Mason J, Peres R, Reynolds D, editors: Proc. Odyssey-04, *The Speaker and Language Recognition Workshop,* 2004.

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

Robertson B, Vignaux GA: *Interpreting Evidence.* Wiley, Chichester, 1995.

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

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

Rose P: Identifying Criminals by their Voice - the Emerging Applied Discipline of Forensic Phonetics. *Australian Language Matters*, 5/2: 6-7, 1997.

Rose P: Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence. *Computer Speech and Language* Special Issue, 20(2-3): 159-191, 2006

Rose P, Lucy D, Osanai T. Linguistic-acoustic Forensic Speaker Identification with Likelihood Ratios from a Multivariate Hierarchial Random Effects Model: A "Non-Idiot's Bayes" Approach. In: Cassidy S. editor. *Proc. 10th Australian Intl. Conf. on Speech Science and Technology.* Sydney: Australian Speech Science & Technology Association, 492-497: 2004

Rose P, Osanai T, Kinoshita Y. Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold. *Speech Language and the Law,* 10/2: 179-202, 2003.

Rose P, Duncan S: Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics* 2/1: 1-17, 1995

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

Dr. Phil Rose, Reader in Phonetics and Chinese Linguistics. School of Language Studies, the Australian National University, Canberra.
philip.rose@anu.edu.au