Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-based Forensic Speaker Discrimination

Phil Rose

Phonetics Laboratory, Faculty of Arts, The Australian National University
philip.rose@anu.edu.au

Abstract

The necessity of taking correlation between variables into account when estimating strength of forensic speaker recognition evidence is argued for. A modest forensic speaker discrimination experiment is described which investigates how well non-contemporaneous speech samples from the same speaker can be discriminated from different-speaker samples using bivariate kernel density likelihood ratios from F2 and F3 of the five monophthongal phonemes of General Australian English, spoken by 11 males. The experiment shows that an approach which takes the correlation of variables into account can yield useful strengths of evidence. It is also pointed out that the results of such tests still require evaluation with the appropriate confidence limits.

1. Introduction

In 1999, Sally Clarke was found guilty by a jury of having murdered her two sons. One was two and a half months, the other two months old at their times of death, in 1996 and 1998 respectively. Clarke was sentenced to life imprisonment. Her case is often cited, eg. [1], as one of the major miscarriages of justice due to egregiously incorrect statistical interpretation. For it was part of the prosecution case that the odds against two SIDS-related deaths in one family similar to hers (the defence case) were 73 million to one. This figure was derived by the prosecution expert as the square of the probability of one SIDS-related death, obtained from relevant literature, of 1 in 8500. As pointed out subsequently by the Royal Statistical Society [2]:

> This approach is, in general, statistically invalid. It would only be valid if SIDS cases arose independently within families, an assumption that would need to be justified empirically. Not only was no such empirical justification provided in the case, but there are very strong a priori reasons for supposing that the assumptions will be false. There may well be unknown and genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely.
>
> The well publicised figure of 1 in 73 million thus has no statistical basis. Its use cannot reasonably be justified as a 'ballpark' figure because the error involved is likely to be very large, and in one particular direction. The true frequency of families with two cases of SIDS may be very much less incriminating than the figure presented to the jury at trial.

In other words, the two deaths were not independent events and the probability of their random conjunction could not thus be estimated from their product according to the third law of probability for independent events. Because of this, and other problems associated with the prosecution case – not the least of which was the failure to disclose crucial information on a possible natural cause of death of the second infant from infection – the case was dismissed on second appeal in 2003 [3].

The Clarke case highlights the serious consequences of the incorrect statistical evaluation of forensic evidence. How does this failure to incorporate dependency of evidence relate to Forensic Speaker Recognition (FSR)? Likelihood ratios have been derived for the comparison of trace evidence (elemental ratios in glass fragments), which take into account correlation between variables, e.g. [4] and [5], but up to now very little attention has been paid to the question of correlated evidence in FSR. Yet it is clear that, given the fact that voices are heavily multidimensional and "... the assumption of independence [of predictor variables] is clearly almost always wrong (naturally occurring covariance matrices are rarely diagonal) ..." [6], the problem needs to be addressed if it assumed, as I do in this paper, that the aim of forensic speaker recognition is to estimate as accurately as possible the LR for the evidence.

Take schwa ([ə]), for instance. According to phonetic theory, a [ə] 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. As is well known, the frequency of a formant from a tract with this articulatory configuration is given by $(2n\text{-}1)^* (C/4l)$, where n is the number of the formant, l is the length of the vocal tract in centimeters, and C is the speed of sound in cms./sec.. This means that all the formants of a [schwa]–sounding vowel must in theory be correlated: if we observe an F1 of 500 Hz in a [schwa] vowel, the acoustic theory of speech production predicts that its F2 will be 1500 Hz and its F3 2500 Hz. It would clearly be wrong to estimate a separate LR for F2 and F3 in [schwa], and then derive an overall LR from their product under the assumption that they were independent variables. Neither is correlation confined to traditional features, like formant centre-frequencies. Automatic features are not immune either: massive correlation has been found between cepstral coefficients in a forensic speaker discrimination experiment [7].

The aim of this paper is to investigate one question concerning correlation in FSR. It asks whether an analytic approach, with a LR formula that takes correlation into account, can usefully discriminate same-speaker from different-speaker speech samples. This question has already been addressed in a large-scale experiment with Japanese, using both traditional and automatic features [8]. However, in that experiment a phonetically heterogeneous and forensically unrealistic set of segments – a long vowel, a voiceless

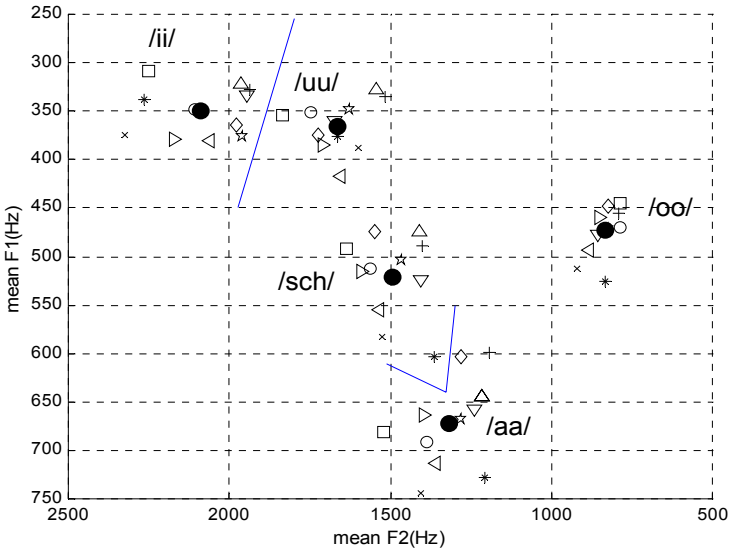

Figure 1: F1 ~ F2 plot of test data vowels, showing vowel means (solid circles) and individuals' means (other symbols). /ii/ = /iː/, /uu/ = /ʉː/, /aa/ = /aː/, /oo/ = /oː/, /sch/ = /əː/.

fricative, and a nasal ([oː ɕ ŋ]) – was used, and it still remains to be seen what performance is like with segments and their features commonly used in real case-work, like the F-pattern in some or all of a speaker's long vowels. Data have recently become available which allow discrimination to be attempted with mean formant centre-frequencies of a set of long vowel phonemes from Australian English, and that is what this paper describes.

As already mentioned, I assume from the outset that the aim of FSR is to estimate a likelihood ratio for the evidence, and not to estimate the probability that a questioned sample was said by a particular person. Given the literature to date on the correct evaluation of evidence in FSR, and many other areas of forensic identification science, e.g. [9], [10], [11], [12], [13], this should now need neither explanation nor justification. Thus the FSR expert has to estimate both the probability of observing the differences between suspect and offender samples assuming they have been said by the same person *and* assuming they have not. "A wise man proportions his belief to the evidence" wrote Hume. The strength of such evidence in favour of the hypothesis that they were said by the same person is the ratio of these probabilities: the more the ratio deviates from unity in either direction, the stronger the corresponding hypothesis. A LR can be used as a discriminant distance – LRs greater than unity being considered as coming from the same speaker, and those less than unity taken as being from different speakers. The degree to which this is correct thus affords a test of the method, e.g. [8], [12], [14]. The testability of the approach is also a vital consideration from the point of view of admissibility of scientific evidence in the wake of the well-known *Daubert* rulings [11].

2. Procedure

2.1 Test data: speakers

Test data were from Alderman's (2005) forensic discrimination experiments [14]. He obtained mean formant centre-frequency data from eleven young(ish) Anglo Australian male speakers (GB ST GK RK ML DS JA AH AC DH TA). The speakers were aged between 18 and 26, and included three pairs of brothers - TA and JA, AH and DH, GK and RK - the last of which were identical twins. All were native speakers of General Australian English (AE), one of the three conventional accent categories for AE (the other two are Broad and Cultivated). The speakers were recorded on two separate occasions separated by at least two weeks. In this way an attempt was made to simulate realistic forensic conditions, at least as far as non-contemporaneity of samples and inclusion of some similar-sounding speakers are concerned.

2.2 Test data: corpus

The corpus consisted of the five long AE monophthongal phonemes /iː/, /ʉː/ /aː/ /oː/ /əː/. Their realisation is implied by their phonemic symbols, except for /iː/, which usually has allophones with centralised on-glides, thus [ə̟iː] or [ɪ̟iː]. The vowels were embedded in /C_d/ words (*heed, deed; hard, card; herd/heard; hoard, board; who'd*), which were elicited in stressed, sentence-final position e.g. *where is the deed?* This was to achieve comparability with the reference data, which also occurred before /d/ in utterance-final position. Twelve replicates were obtained in both of the recording sessions. No particular attempt was made to control for recording venue or mike. The vowel acoustics (F1, F2, F3) were measured at target using *Praat*. The speakers' vowel statistics are summarised on pages 52-57 of [14]; individual observations are given in appendix 3 of [14].

Figure 1 shows a conventional F1 ~ F2 plot of the eleven speakers' test data. To avoid cluttering, each individual's first and second recording session means were pooled, and thus they represent the means of 24 replicates. The mean vowel positions in the F1/F2 plane are typical of General AE, and reflect the allophones described above, with the exception of /ʉː/, which has probably been pulled forward of central by the following alveolar consonant. The typical configuration can be seen of high, mid and low central vowels (/ʉː/ /əː/ /aː/), with a single non-low vowel on either side.

2.3 Reference data

Estimation of a LR requires a reference, or background, distribution, against which the typicality of the pair of samples being tested can be assessed (a LR is a ratio of similarity to typicality). The Bernard data set was used. This is a set of formant values (F1, F2 F3) from the vowels of some 170 Australian English males collected in the 1960's [15]. Bernard assigned his subjects to one of the three conventional Broad, General or Cultivated accent categories of AE, and recorded them saying vowels in /h_d/ words: once with the word in isolation ("CIT") and once with the word in stressed sentence-final position ("FRAME"). An attempt was also made to elicit prolonged vowel tokens, but many speakers were unable to do this satisfactorily. The Bernard data – summaries and all

individual observations – are provided in appendix 2 of [14].

Table 1: Within- and between-vowel correlations (Spearman's *rho*) between F2 & F3 in Bernard reference data.

	/aː/ F2	/aː/ F3	/iː/ F2	/iː/ F3	/ʉː/ F2	/ʉː/ F3	/əː/ F2	/əː/ F3	/oː/ F2	/oː/ F3
/aː/ F2										
/aː/ F3	**0.119**									
/iː/ F2	0.165	0.034								
/iː/ F3	0.086	0.210	**0.525**							
/ʉː/ F2	0.310	0.106	0.009	0.082						
/ʉː/ F3	0.403	0.149	0.173	0.084	**0.379**					
/əː/ F2	0.500	0.010	0.165	0.131	0.299	0.362				
/əː/ F3	0.395	0.231	0.318	0.348	0.313	0.471	**0.377**			
/oː/ F2	0.187	0.281	-.128	0.089	0.022	0.241	0.105	0.015		
/oː/ F3	-.178	0.123	0.262	-.001	-.056	0.032	-.296	0.209	**-.115**	

Since the test speakers in the present experiment are described as having a General Australian accent, and since their data occurred in stressed sentence-final position, data from the FRAME subset of the General set of Bernard speakers were used as reference. Tokens were available for about 60 speakers, except for F3 in /oː/, where, not surprisingly, Bernard was able to make measurements of only 38. In size, the Bernard data set, and its subset of General speakers, must be considered a reasonably representative reference sample. It is also nicely comparable with the test subjects, being homogeneous with respect to ethnicity and age (Bernard's subjects were mostly Anglo university students). However, as it is separated from the test data by some thirty-five years, and as there have been changes in AE accents in this time, there are naturally questions as to its representativeness in age. It is therefore important to note that the differences in formant centre-frequency between the Bernard data set and more modern AE accents have been shown not to be large enough to affect the former's use as reference sample in forensic discrimination [16] and [17].

3. Correlation

Since this paper is about correlation between variables, it is important to examine the actual correlation structure of the data. This is shown in table 1, which contains the correlation matrix (Spearman's *rho*) for the vowels' F2 and F3 in the Bernard General AE combined FRAME and CIT data. Thus a typical correlation calculation was based on two values for each speaker for each vowel formant: about 120 pairs. For this degree of freedom, it must be remembered that the significance level threshold, even at the 99% confidence limit, is quite low: for df = 120 it is 0.23; for df = 60 it is 0.325. Thus even low values for *rho* indicate the likely presence of correlation.

Table 1 shows that, as expected, correlations are present in the data, although it can be seen that, as far as within-vowel correlations are concerned, F2 and F3 are not necessarily correlated for each vowel. They are essentially uncorrelated in /aː/ and /oː/; show mild correlation in /əː/ and /ʉː/; and are fairly strongly correlated in /iː/. As far as between-vowel correlations are concerned, most are very low, although mild correlation can be seen

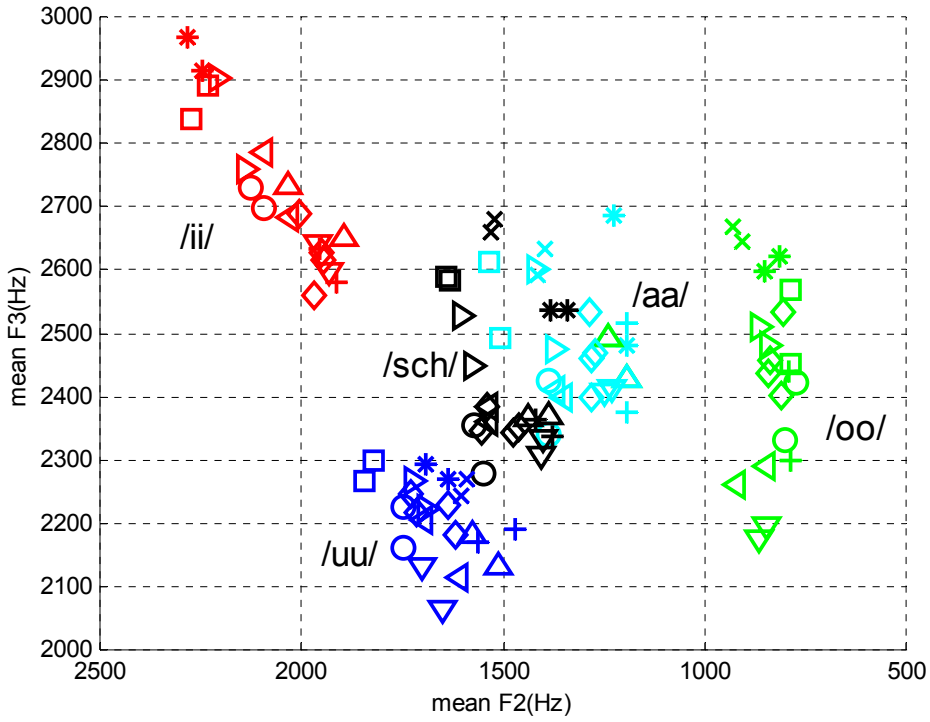

Figure 2: F2 ~ F3 plot of test data vowels, showing 11 speakers' means from both recordings. Each speaker is indicated with a different symbol. /ii/ = /iː/, /uu/ = /ʉː/, /aa/ = /aː/, /oo/ = /oː/, /sch/ =/əː/.

between some central vowel formants. For example, F2 in /aː/ correlates with formants in the other two central vowels /ɐː/ and /əː/; F3 in /əː/ shows correlations with all vowels, especially F3 in /ɐː/. It is therefore legitimate to ask how these correlations should be handled.

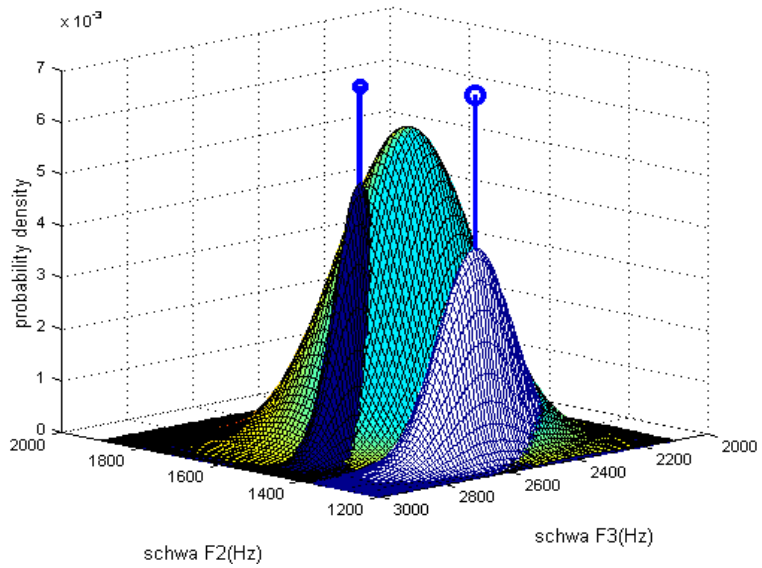

Figure 3: Joint bivariate normal probability density distributions for /əː/ F2 and F3 in Bernard's reference population, and in samples from identical twins (GK, RK). Vertical lines show location of means from GK (large ball on top) and RK (small ball).

4. Processing

Telephone transmission affects all formants, but in real case-work the first formant is often badly compromised, which can then be exacerbated by poor automatic extraction. Because of this, only the (relatively high) F1 of low and mid vowels, e.g. /aː/ and /əː/, are candidates for use with telephone recordings, and even then it is difficult to be totally sure that they remain uncompromised. In order to conform to realism, then, it was decided in the main experiment to only use the vowels' F2 and F3 measurements, even though telephone recordings were not used. The inclusion of F3 can be justified by the fact that it has been shown to perform well in realistic forensic discrimination, even between non-contemporaneous *telephone* and *direct* recordings [10, pp. 5107-5112], see also fig. 3, p. 176 of [11]. In these cases, transmission degradation was obviously not great enough to adversely effect the discrimination, and in fact the results in figure 2 show that is only in /iː/ that F3 approaches the nominal upper transmission cut-off of 3.5 kHz. In real case-work, of course, which formants are included and which excluded depends on the assessment of the transmission degradation (from long-term spectra, for example), and the segment under comparison (even F4 might be sufficiently low in some rhotics). One cannot automatically assume that F3 is inviolate, just as one cannot automatically assume that F1 in /aː/ and /əː/ is unreliable.

The raw mean data available for comparison is shown in a conventional F2 ~ F3 plot in figure 2. This time, the mean values for each speaker's two recording sessions are plotted separately, to allow the reader to appreciate the variation as a function of non-contemporaneity. These are the data to be discriminated, and they are typical. There are thus many cases where the values of the same speaker's two recording sessions are quite close, and different speakers' data are fairly well separated. Nevertheless there are also instances of the same speaker's non-contemporaneous data not being similar; and data from different speakers' recordings being very similar. Note how the plot also shows a clear positive correlation between F2 and F3 in /iː/, and lack of correlation in /oː/, as with the reference data.

A likelihood ratio-based discrimination was then run, for each of the five vowels, treating the speakers' separate sessions as 'suspect' and 'offender' samples to be evaluated against the corresponding Bernard reference distribution. Each vowel was characterised by its twin variables of mean F2 and F3, and a LR derived which took the degree of correlation between variables into account. This is represented graphically in the 3-d plot in figure 3. This depicts a different-speaker comparison, using /əː/, between the first session data of the two identical twins GK and RK. The joint bivariate probability density distribution for F2 and F3 in /əː/ for the 60 speakers in the Bernard reference data is the large mound in the middle of the plot. Most of RK's distribution, similar in F2 to the reference distribution, but lower in F3, protrudes from its middle. GK's distribution, similar to the reference in F3, but considerably lower in F2, lies to its bottom right. GK's and RK's bivariate /əː/ distributions are well separated in the distributional projection of the F2~F3 plane, and are also fairly atypical of the reference distribution. It is no surprise, therefore, that the LR for their comparison, as estimated from the formula to be used in this paper, is astronomical: 1.4 E-14. This means that one would be far, far more likely to get this difference in /əː/ F2 and F3, given the reference distribution, had the two samples come from different speakers. Since they have, that is a good result. Identical twins, obviously, do not necessarily have identical mean F-pattern values.

The enormous magnitude of the LR for this particular comparison should not blind us to the fact that there are plenty of different-speaker comparisons that yield LRs greater than unity - that is, the difference between the samples has to be interpreted as more likely had they come from the same speaker - and there are also plenty of same-speaker comparisons with LR < 1. This is the way speakers and their formants behave.

The formula for the LR used in this paper was developed as a solution to the non-trivial problem of estimating the strength of evidence when predictor variables may be correlated. It treats the variables for which a LR has to be

numerator of MVLR = (1)

$$(2\pi)^{-p}|D_1|^{-1/2}|D_2|^{-1/2}|C|^{-1/2}\left(mh^p\right)^{-1}\left|D_1^{-1}+D_2^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}$$
$$\times\exp\left\{-\frac{1}{2}\left(\bar{y}_1-\bar{y}_2\right)^{\mathrm{T}}\left(D_1+D_2\right)^{-1}\left(\bar{y}_1-\bar{y}_2\right)\right\}$$
$$\times\sum_{i=1}^{m}\exp\left[-\frac{1}{2}\left(y*-\bar{x}_i\right)^{\mathrm{T}}\left\{\left(D_1^{-1}+D_2^{-1}\right)^{-1}+\left(h^2C\right)\right\}^{-1}\left(y*-\bar{x}_i\right)\right]$$

denominator of MVLR = (2)

$$(2\pi)^{-p}|C|^{-1}\left(mh^p\right)^{-2}\prod_{l=1}^{2}\left[|D_l|^{-1/2}\left|D_l^{-1}+\left(h^2C\right)^{-1}\right|^{-1/2}\times\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}\left(\bar{y}_l-\bar{x}_i\right)^{\mathrm{T}}\left(D_l+h^2C\right)^{-1}\left(\bar{y}_l-\bar{x}_i\right)\right\}\right]$$

where U, C = within-, between-speaker variance/covariance matrices; n_1, n_2 = number of replicates per speaker
m = number of speakers in reference population; p = number of assumed correlated variables per speaker
$D_l = D_1, D_2$ = offender, suspect var/cov matrices = $n_1^{-1}U, n_2^{-1}U$
h = optimal smoothing parameter for kernel density = $\left(4/(2p+1)\right)^{1/(p+4)}m^{-1/(p+4)}$
$\bar{y}_l = \bar{y}_1, \bar{y}_2$ = offender, suspect means; $y^* = \left(D_1^{-1}+D_2^{-1}\right)^{-1}\left(D_1^{-1}\bar{y}_1+D_2^{-1}\bar{y}_2\right)$
$\bar{x}_i$ = within-speaker means of reference population.

estimated – for example a set of measurements of frequencies for different vowel formants – as multivariate data [4], and hence its output is called a multivariate LR (MVLR). The between-group variance can either be considered normal or estimated with a kernel density. Inspection of the reference formants' univariate distributions showed clear deviations from normality in some cases, and it was therefore decided to extract kernel density LRs. The approach accommodates two levels of variance: between- and within-subjects, and is thus somewhat unrealistic for speech, where at least a third level of variance – between non-contemporaneous sessions – must usually be assumed, since the within-speaker variance between sessions is usually greater than within a session [9], [19]. The approach is perhaps also unrealistic for speech in requiring the within-subject variance to be modeled normally. The formula for the multivariate kernel density LR, from pp.116,117 of [4], is reproduced at (1) and (2). (1) is the expression for the numerator, (2) the denominator. The numerator can be seen to quantify the similarity between the mean values of the offender and suspect; the denominator quantifies the typicality of the difference against the reference population. The LR is the ratio of their values.

Bivariate MVLRs were calculated for each vowel separately. This yielded, for each of the five vowels, a list of 11 LRs from the same-speaker comparisons, and 110 LRs from the different-speaker comparisons. Of course, it is natural to also want to know how the discrimination performs with various sub-sets of vowels, for example, all vowels combined. Here, however, a multivariate LR cannot be used, since there is no sense in which, say, an individual F2 observation from one token of a speaker's /a:/ vowel is multivariate with an F2 observation in a token from a different vowel, say /i:/. In order to get an idea of how various vowel combinations perform, then, independence LRs were estimated for three sub-sets of the data: for all vowels combined; for the central vowels; and the optimum set of vowels. In other words, overall LRs were estimated by taking the product of the individual vowels' bivariate LRs, under the assumption they are independent. As already demonstrated (see table 1), this assumption is not always correct. However, in order to legitimately treat formant data from different vowels as multivariate, one would need mean data from many more sessions. This is a job for the future.

5. Results

The result of the discrimination was, for each vowel, a list of 11 LRs from the same-speaker comparisons, and 110 LRs from the different comparisons. The LRs were then plotted using a so-called Tippett display. Tippett plots are cumulative distributions of LRs from same-subject and different-subject trials. They show for what proportion of same- or different-speaker trials one observes a LR bigger than a given abscissa LR value, and are a popular way of displaying forensic discrimination data [12].

Normally in Tippett plots, the same-speaker LRs are plotted inversely, but this makes it difficult to see details like the EER, and is avoided here. Also, LRs for different-speaker comparisons can get very big – witness the LR for the identical twin comparison above. This is because, within any sample, some different-subject pairs will differ more than others. Also, at least for traditional features, same-speaker LRs do not typically get anywhere near as big as different-speaker LRs. This is because two samples cannot get more similar for a feature than identical, and under these circumstances, other terms in the LR formula, like the number of items in a sample, and especially the ratio of within- to between-speaker variance for the feature, have a limiting effect on the magnitude of the LR. Realistically, therefore, there is little sense in plotting the whole of the LR range for discrimination, and it is better to concentrate on the details near threshold, say in the range between LogLR values of -5 and 5.

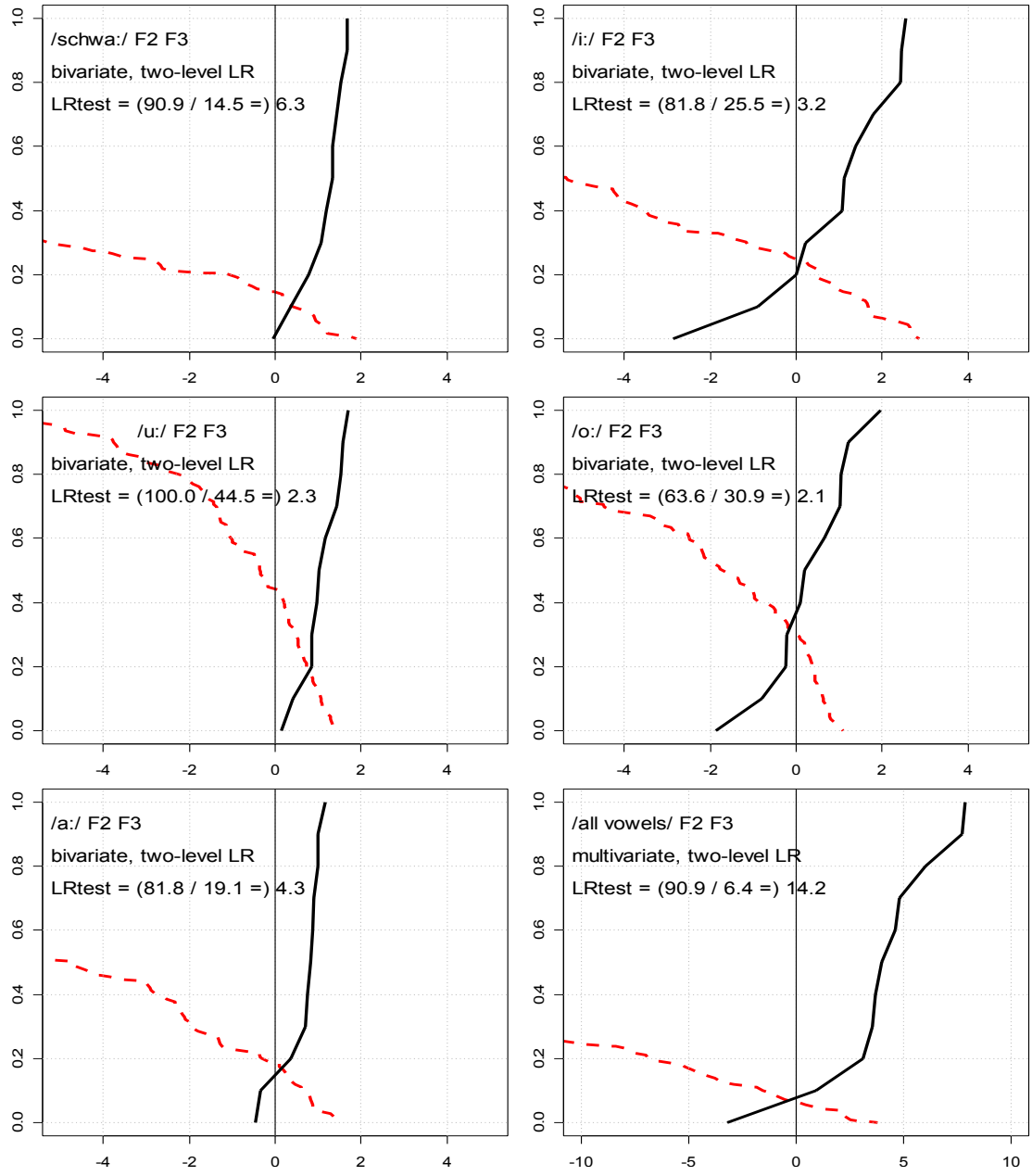

Figure 4: Tippett plots for bivariate kernel density LR discrimination with F2 and F3 of the five separate Australian long vowels: /schwa: i: a: o: u:/, and all vowels combined (bottom right).Dotted line = LRs for different-speaker pairs; solid line = LRs for same-speaker pairs. Horizontal axis = logLR, vertical axis = cumulative probability. Vertical line = threshold.

Figure 4 shows the Tippett plots for each of the five vowels, and for all five vowels combined, based on bivariate LRs from correlated F2 and F3. Taking /əː/ as an example (in the top left panel), it can be seen that the cumulative plot for its same-speaker LRs lies almost totally above the threshold (of LogLR = 0), indicating that nearly 100% of the same-speaker pairs had, as hoped, LRs greater than threshold when compared using F2 and F3 in /əː/. In reality, 10 out of the 11 same-speaker pairs were correctly discriminated (this is shown by the figure of 90.9 in the LR*test* caption). This indicates that if one were to get a LR less than threshold for a bivariate comparison using /əː/ F2 and F3, there would be about a 10% chance of error – that the same speaker were involved rather than different speakers.

The different-speaker LR line crosses threshold at about 15%, which reflects the fact that, of the 110 different-speaker pairs tested, 16, or 14.5%, were incorrectly evaluated. Thus, given a LR greater than threshold, the risk of a judicially fatal

error (saying that the same speaker is involved when in reality they are not) is a rather high 15%.

It can also be appreciated that the figure shows only a small number of the different-speaker LRs: 70% of them have LogLRs less than -5. This magnitude is not symmetrical: all same-speaker LRs for /əː/ have values below 100. The EER, of about 10%, can also be seen at a value slightly greater than the 0 LogLR threshold, where probability theory predicts it to fall. (The threshold of 0 LogLR in LR testing is sacrosanct, because it is required by theory [11].)

The LR for the test ("LRtest") is 6.3. This indicates that, with a bivariate LR from /əː/ F2 and F3, one would be about six times more likely to get a LR greater than threshold assuming that the samples came from the same rather than from different speakers. In the verbal scale proposed for the UK Forensic Science Service [18], this would be characterised as only "limited" evidence (in support of the hypothesis that the samples came from the same speaker). The Tippett plot shows that this strength of evidence would increase to barely "moderate" for LogLRs greater than 1, since one would be about [p(LogLR > 1 | SS) / p(LogLR > 1 | DS)] =13 times more likely to get a LogLR greater than 1 assuming same rather than different speakers, and "moderate" is the term used to characterise the strength of such values (i.e. between LogLR = 1 and LogLR = 2).

Figure 4 shows that all the individual vowels have some discriminatory potential and evidential value, with /əː/ and /aː/ being the strongest in terms of LR*test*. However, the strength of the evidence from any one vowel is still, at the best, only moderate. When the bivariate LRs from all vowels are combined (in the bottom right panel), there is considerable improvement. It can be seen firstly, and most importantly, that the risk of a judicially fatal error [p(SS|DS, LogLR > 0)] has fallen to about 6%. There is also an increase in the LR*test*. One would, with all vowels combined, be about 14 times more likely to get a LogLR greater than 0 if the data were same-speaker. With a LogLR of 1000 or more, this increases to about 80 times more likely, and, although still classified as "moderate", is considerably better than the value of 13 quoted for /əː/. That one can contemplate same-speaker LRs of 1000 or more is of course because the same-speaker LRs are distributed considerably higher: note the much higher upper bound for same-speaker LRs, at about 32 million. (Although one has to be cautious here, since these higher values may in part be due to illegitimate combination of between-vowel correlated evidence!) Finally, the EER is now at about 7%, and located very close to the LogLR threshold of 0, nicely in accordance with probability theory.

As mentioned above, discriminations were also conducted with subsets of the five vowels. Excluding /iː/ actually gave the best discriminatory performance of the whole experiment, with a LR*test* of 33 (EER ≈ 3%), where the good performance is mainly due to the good resolution of different-speaker pairs, with only 3 out of 110 comparisons incorrect. This is shown in figure 5. The only combination to yield 100% correct same-speaker performance was with the natural class of central vowels /ə aː uː/. This combination still misclassified 9.1% of different-speaker pairs, and had a LR*test* of 11. This good same-speaker performance may reflect an overall lower within-speaker variance for central vowels, but it is always difficult to account for differential performances of subgroups

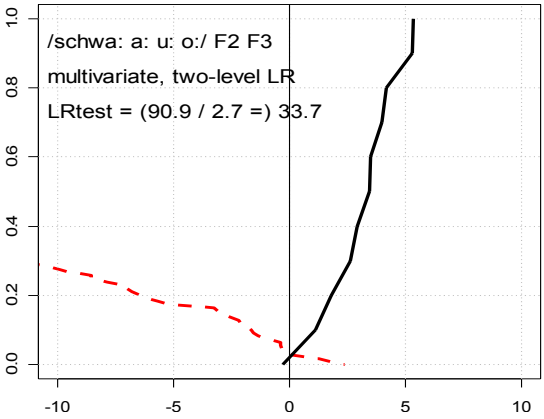

Figure 5: Tippett plots for bivariate kernel density LR discrimination with F2 and F3 of the four separate Australian long vowels: /schwa: aː oː uː/.

in an experiment like this, where the outcome is the result of a complex interplay of many variables.

Finally, the effect of including F1 for the vowels /aː/ and /əː/ was investigated. This was done by estimating a multivariate LR for all three formants, and then comparing the result with the corresponding bivariate LRs. It was found that the discriminatory potential was not affected, but the LR was shifted further away from threshold by a factor of 2 to 3 for same-speaker comparisons, and by a factor of 1.3 to 5 for different-speaker comparisons. This shows that the F1 of /aː/ and /əː/ contains speaker-specific information that can contribute to a LR-based comparison, and clearly it is an advantage to include it, as long as the investigator can be sure that the F1 has not been affected by the telephone transmission.

6. Confidence limits for Tippett plots

One question about Tippett plots that does not yet seem to have been asked concerns their confidence limits. For example, using F2 and F3 in all vowels, 103 out of the 110 different-speaker pairs were correctly evaluated. Modeling these data with the appropriate beta(α, β) distribution, as recommended by the *European Network of Forensic Institutes* (p.185 of [1]), shows that you could be 95% sure that the actual discrimination was at least 89.5%. The considerable decrease in the number of same-speaker comparisons (11) of course brings with it a much lower confidence limit: even though the correct same-speaker pair discrimination was 90.9%, the 95% lower limit for that number of comparisons is actually 74%. Since the main function of Tippett plots is to convey to the court the reliability of the method [12], it is appropriate that their associated confidence limits also be estimated as part of that information. In that way they can then be legitimately construed as answering the 'pre-data' question: how likely am I to make a mistake with this approach? [4].

7. Summary

This paper was motivated by the fact that an accurate estimation of a LR in forensic speaker recognition must involve taking any correlation between variables into account.

It has shown that correlations exist between F2 and F3, both within and between vowels, and has investigated the use of bivariate LRs to properly account for such within-vowel correlation. A LR-based forensic speaker discrimination using bivariate LRs was used to quantify the discriminant performance of the five long monophthongal vowel phonemes of eleven male speakers of General Australian English, both individually and combined. It was shown that each vowel had at least some discriminatory potential, and that all five combined yielded strengths of evidence that could be characterised as at least "moderate". These results show once more that it is possible to discriminate same- from different-speaker pairs using traditional acoustic features and a LR as a discriminant distance.

The discriminant performance, with EERs of 7% or so, is not stellar. It does, however, give us an idea of what strengths of evidence are to be expected from real-world forensic comparison using mean formant centre-frequencies. Most importantly, the results further suggest that the approach will be of use in real case-work, where it is quite often possible to quantify and compare vowels with respect to F2 and F3. Indeed, given the correlations demonstrated, it will be mandatory for all vowels except /a:/ and /o:/, should the investigator want to compare offender and suspect samples with respect to both F2 and F3 in a given vowel.

There are several next steps. One is to compare the present results with those derived with an "independence Bayes" approach. This will quantify the effect on LRs of taking correlation into account. Another is to derive a more appropriate LR formula for speech which is able to take three levels of variance into account. And lastly, enough data must be gathered to investigate between-vowel correlations in F-pattern: it must not be forgotten that the effect of such correlations on LRs still remains unaccounted for.

8. Acknowledgements

I should like to thank Tony Alderman, formerly of the Australian National University's *National Dictionary Centre*, for letting me use his formant data for re-analysis. Now a member of the Australian Crime Commission, I hope he does not also become forever lost to forensic speaker recognition. I also want to thank Dr. David Lucy of Edinburgh University's *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* for making available the original multivariate LR programs in R. Thanks also to my two reviewers for their comments.

9. References

[1] Lucy, D., *Introduction to Statistics for Forensic Scientists*, John Wiley, Chichester: 159 – 160, 2005.

[2] RSS www.rss.org.uk 23rd Oct. 2001.

[3] Johnson, P., "The Sally Clarke Case: Another Collision Between Science and the Criminal Law", *Australian Journal of Forensic Sciences*, 36:11-33, 2004.

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

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

[6] Hand, D.J. and Yu Keming, "Idiot's Bayes – Not So Stupid After All?", *International Statistical Review*, 69(3):385–398, 2001.

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

[8] Rose, P., Osanai, T. and 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.

[9] Rose, P., *Forensic Speaker Identification,* Taylor and Francis, London & New York, 2003.

[10] Rose, P., *The Technical Comparison of Forensic Voice Samples*, Issue 99, Expert Evidence, Freckelton, I. and Selby, H., series eds., Thomson Lawbook Company, Sydney, 2003.

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

[12] Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M. and 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.

[13] Aitken, C.G.G., and Taroni, F., *Statistics and the Evaluation of Evidence for Forensic Scientists*, Wiley, Chichester, 2004.

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

[15] Bernard, J.R.L., Some measurements of some sounds of Australian English, unpublished Ph.D. Thesis, university of Sydney, 1967.

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

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

[18] Champod, C. and Evett, I., Commentary on [19], *Forensic Linguistics* 2000, 7(2): 238-43, 2000.

[19] Broeders, A.P.A., "Some observations on the use of probability scales in forensic identification". *Forensic Linguistics*, 6(2): 228-41, 1999.

[20] Rose, P., "Long- and Short-term within-speaker differences in the formants of Australian *hello*". *Journal of the International Phonetics Association*, 29/1: 1-31, 1999.