The Bernard Data Set as a Reference Distribution for Bayesian Likelihood Ratio-based Forensic Speaker Identification using Formants.

Tony Alderman
Phonetics Lab (Arts) & Australian National Dictionary Centre
Australian National University,
alderman@webone.com.au

Abstract

This paper examines the use of John Bernard's data set of male speakers of Australian English as a reference distribution for the practice of Forensic Speaker Identification in Australian contexts. The Bayesian likelihood ratio-based discrimination is performed on recordings of eleven male speakers of Australian English, using the f-patterns of F1, F2, and F3 of their five tense monophthongs /i/, /a/, /o/, /ʉ/, and /ɜ/. Target values for these formants and vowels of the speakers described as Broad or General by Bernard are used as the reference distribution for the analysis. The performance of the data set is evaluated using a Likelihood Ratio as discriminant function, and also by an examining of the strength of evidence generated. The analysis is performed using a relatively simple model, assuming normality of the reference sample and equal variance. It is shown that the Bernard data function well, yielding strong strength of evidence. The results of the analysis are discussed with regard to the practice of FSI in Australia.

1. Introduction

The application of Forensic Speaker Identification (FSI) in the real world typically involves comparison of two speech samples. The data for this comparison usually comprises samples of an unknown speaker, generally the alleged offender, and another known speaker, generally the suspect. This data may consist of recordings of telephone calls, police interview tapes, surveillance videos and the like. The practitioner of FSI is usually asked to make some kind of statement regarding the similarities/differences between the two voices. The proper evaluation of these similarities/differences in FSI is optimally performed using a Bayesian framework, with the evaluation of evidence presented in terms of a Likelihood Ratio (Rose, 2002, 2003; Champod and Meuwly, 2000; Aitken, 1995).

The Likelihood Ratio is a measure of the strength of evidence supporting one of two competing hypotheses. This is not the same as providing a statement of the probability of the hypothesis given the evidence; this requires access to prior odds not generally available to the analyst. It is further not the place of the analyst to provide such a statement in a legal context, as this is the proper domain of the judiciary (Robertson and Vignaux 1995). The LR takes the form of a number expressing the ratio of the probability of evidence assuming one hypothesis, divided by the probability of evidence assuming the competing hypothesis. This is shown at (1), where *p* represents probability, H a hypothesis of identity, H_A a hypothesis in competition with H, and E represents the evidence under consideration.

$$LR = \frac{p(E|H)}{p(E|H_A)} \quad (1)$$

In FSI, H is typically the prosecution hypothesis that the samples have been produced by the same speaker, and H_A the defence hypothesis that the samples were produced by different speakers. A LR greater than 1 thus indicates a higher relative probability of the evidence given the prosecution hypothesis, and a value less than one the opposite. The magnitude of the distance of the LR from 1 (unity) indicates how strongly the evidence supports one of the two hypotheses.

Verbal equivalents of the strength of the LR often use Log_{10} as a base, such that a LR of 30 is not considered any stronger than a LR of 14 – it is at 100 that the scale moves from moderate to strong support (González-Rodríguez, Ortega-García & Sánchez-Bote, 2002: 174). In such a Log based representation of the LR, the threshold for discrimination also of course moves from 1 to 0, with different-speaker LRs expected to be negative values instead of fractions.

In order to accurately assess the similarity of the two voices, as well as evaluate the strength of evidence supporting the different-speaker hypothesis, a sample representing the population must be produced, in order to evaluate just how typical of the different speakers in question the samples being analysed are. Two samples may show remarkable similarity to each other, but this in itself is not sufficient to support the hypothesis that they were produced by the same speaker. It must also be

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

evaluated just how typical this feature is in the relevant population. If it is very common to have such a feature in the population the similarities between the samples become less supportive of the same-speaker hypothesis.

A problem in the application of FSI in Australian contexts, indeed in most contexts, is the scarcity of adequate background populations for evaluation of similarity and typicality in acoustic forensic phonetic parameters. In the late 1960s John Bernard collected and analysed the speech of 170 male speakers of Australian English (AE), in terms of the first three formants for the vowels of AE (Bernard, 1970, 1989). The speakers were also categorised according to one of the three accent types proposed by Mitchell and Delbridge (Mitchell and Delbridge 1965). This means that the data set is potentially useful for a number of different variations of the different-speaker hypothesis based on accent descriptions (such as different speaker, different broad speaker, etc.).

The testing of the data set provides an empirical foundation for the use of the data set in real forensic situations, in accordance with criteria of scientific robustness necessary after *Daubert* for scientific evidence presented in a forensic context (Black, Francisco, and Saffran-Brinks, 1994).

The first aim of the experiment in this paper is therefore to test the Bernard data set to evaluate its use in the actual application of FSI in Australia. This is performed using target values for the first three formants of the /tense monophthongs/ of AE as recorded by Bernard.

The second aim of the experiment is to test the performance of the F-patterns for the first three formants of the /tense monophthongs/ for speaker discrimination. These aims are carried out using speech from 11 male speakers of AE, aged between 18 and 26.

In order to evaluate the performance of the set, both magnitude of LRs and also the LR as a discriminant function are examined.

This paper will show that not only is the Bernard data set usable for FSI of male speakers of AE, but that the distribution of results is of a type which is useful in legal contexts, in that different-speaker pairs are discriminated much better than same-speaker. This means that while the frequency of same-speaker determinations is somewhat low, there are very few false positives. In practice, this means that it is less likely to obtain a result which would provide support (erroneously) that same-speaker production of the speech samples is involved, and thus perhaps contribute to an innocent person being found guilty by the court.

2. Procedure

The experiment uses data collected for the experiments conducted in Alderman (2004). This comprises recordings of eleven male speakers of AE, aged between 18 and 26. Two of the speakers were identical twins. Tokens of the tense monophthongs were collected in a controlled environment, with the vowels elicited in a /h_d/ context, in a stressed sentence-final position (such as “that wasn’t very hard”, for /a/). Whilst this method of elicitation has some shortcomings in conforming to the criterion for optimal speech samples used in FSI analysis (Rose, 2002), it does provide a nicely comparable structure to the Bernard data, in controlling for the effect on F-pattern of the perivocalic segments, which is useful at this stage of determination of suitability of the Bernard set for FSI.

Two recording sessions for each speaker were conducted, with twelve tokens of each vowel elicited in each recording session. The recordings were separated by at least two weeks, to introduce the necessary within-speaker variation which is a feature of naturally occurring speech, and also a crucial desideratum for FSI experiments of this kind (Rose, 2002).

2.1. The Normal Approach

There is more than one method for calculation of the LR using continuous acoustic parameters such as formants. Both have strengths and limitations, and this varies depending on the structure of the data available for use as a background sample. The first of these assumes normality in the distributions of the variables, and makes use of the mean and standard deviation of the sample. There are a number of variants of this formula, but Lindley (1977) derives one such version, which also – rather unrealistically for speech - assumes equal variance for both samples. This formula is shown at (2).

$$\mathrm{LR} \approx \frac{\tau}{\mathrm{a}\sigma} \times \underbrace{e^{\left\{-\frac{\left(\bar{x}-\bar{y}\right)^2}{2\mathrm{a}^2\sigma^2}\right\}}}_{\text{similarity term}} \times \underbrace{e^{\left\{-\frac{(w-\mu)^2}{2\tau^2}+\frac{(z-\mu)^2}{\tau^2}\right\}}}_{\text{typicality term}}$$

$\bar{x}$ = mean of questioned sample; $\bar{y}$ = mean of suspect sample

μ = mean of reference sample

σ = standard deviation of questioned and suspect samples

τ = standard deviation of reference sample

$z = \left(\bar{x}+\bar{y}\right)/2$

$w = \left(m\bar{x}+n\bar{y}\right)/\left(m+n\right)$

m = number in questioned sample

n = number in suspect sample

$\mathrm{a} = \sqrt{1/m+1/n}$

(2)

The second method of estimating a LR – by kernel density - is capable of modelling non-normal background distributions, but requires accurate

Accepted after full paper review

estimation of within-speaker variance for a reliable LR calculation. The Bernard data comprises only two or three tokens for any given vowel for any given speaker. This results in poor estimation of the within-speaker variance. Although Alderman (2004) found this method promising, the difficulty in accurate estimation of the within-speaker variance remains a problem.

Additionally, the formula given at (2) was found in Alderman (2004) to provide better discrimination for some combinations of parameters. The formula given at (2) was used in the calculation of the LRs for this experiment. This model does still assume independence of the parameters, which has not been established for this data. This is because whilst theoretically this may bias the generated LR, this is an empirical test to see which formants, and combinations of formants provide the best performance for the practice of FSI in Australia. It was thought useful to at least examine the results of such combinations.

2.2. The Experiment

The two non-contemporaneous recordings of the eleven speakers give eleven same-speaker comparisons and 220 different-speaker comparisons. LRs were calculated for F1, F2, and F3 of each of the tense monophthongal phonemes /i/, /a/, /o/, /ʉ/, and /ɜ/, providing 15 individual LRs for each speaker pairing for analysis. Combinations of LRs were calculated to find the best performance in terms of discriminating between same-speaker and different-speaker pairs. The LRs for those combinations of parameters found to perform speaker discrimination most successfully were then examined in terms of the magnitude of support they provided for the hypothesis.

For the purposes of this experiment, based on arguments and observation of variation in realised accent by social context and other factors (Horvath, 1985; Cox and Palethorpe 1998), a combination of the data from the Broad and General speakers of the Bernard data set was used. Different combinations are possible based on different defence hypothesis formulations, but the Broad and General combination was found to be most representative of the eleven speakers recorded for the test, and also to provide optimal results compared with other combinations. This provides 118 speakers with 2 or 3 tokens per speaker (this varies by vowel) for the reference distribution.

3. Results

Table 1 presents the results of the experiment for the individual LR calculations, for F1. The vowel is listed in the top row. The figure in the SS row is the percentage of correct same-speaker comparisons out of the eleven same-speaker pairings. The DS row lists the percentage of correct different-speaker comparisons out of the 220 different-speaker pairings. The LR_{test} is a crude way of quantifying the strength of evidence (the results) supporting the hypothesis that a same-speaker pairing will be resolved with a LR greater than 1, compared to the hypothesis that a different-speaker pairing will be resolved with a LR>1. It is thus a ratio of the conditional probabilities $p(LR>1 \mid SS)/p(LR>1 \mid DS)$ - the number of correctly discriminated same-speaker pairings to the number of incorrectly discriminated different-speaker pairings. The LR_{test} values have been rounded to 1 decimal place.

Table 1: LR results for F1

F1	/i/	/a/	/o/	/ʉ/	/ɜ/
SS	45.45%	72.73%	45.45%	36.36%	63.64%
DS	67.73%	75.91%	70.91%	66.36%	76.82%
LR_{test}	1.4	3.0	1.6	1.1	2.7

Thus, it can be seen that for F1 in /i/, 45.45% of the same-speaker comparisons (5 out of 11) were correctly discriminated with a LR>1, whilst 67.73% (149) of the 220 different-speaker pairings were correctly discriminated with a LR<1. This gives a LR_{test} value of (45.45/(100-67.73)) 1.4. One would be therefore 1.4 times more likely to get a LR bigger than unity for F1 of /i/ assuming that the pairing actually is the same-speaker, than if they were different speakers. This would of course constitute effectively useless evidence, since it shows that one is almost as likely to get the difference between the samples assuming that they were produced by the same speaker as assuming they were produced by different speakers.

The results for F1 show only limited same-speaker discrimination, with different-speaker pairings being better resolved than same-speaker for F1 of any of the five vowels. This is fairly typical for tests using such a model. This may have to do with limits on identity, whilst difference is theoretically less limited; that is, while things can only ever get as close to identical as identical, the degree of difference between samples can be much greater.

For the five vowels tested, for F1, /a/ and /ɜ/ exhibited the highest correct resolution of both same- and different-speaker pairings. This is just below 73% for /a/ for same-speaker comparisons, and nearly 76% for different-speaker. This results in a LR_{test} of 3 – one would be three times more likely to get the observed difference between the samples assuming same-speaker provenance. /ɜ/ performs nearly as well for SS pairings (63.64%), and slightly outperforms /a/ for the DS pairings (with 76.82%). This gives us an LR_{test} for /ɜ/ of 2.7.

These are good results, as, of vocalic F1, it is values of /a/ and /ɜ/ that are most likely to be useful in FSI in

Accepted after full paper review

the real world. This relates in part to the fact that much of the data used in FSI comparisons comprises recordings of intercepted telephone conversations. The use of telephone recordings for FSI can be problematic, due to interference, poor line quality, and importantly, the band-pass filter systems used in telephone systems (Künzel, 2001: 87-89, 93-96; Rose, 2003: 5101-5113).

This filter effectively distorts information occurring below ca. 500 Hz and above ca. 3 kHz, meaning that for much data F1 (which for a male often falls below this threshold) is, often, an unreliable parameter to use for analysis. /a/ and /ɜ/, out of the five vowels tested, display the highest F1 values, typically above this 500 Hz threshold, meaning that not only do they perform best of the F1s in this empirical examination, but they are also the most likely to be usable in the real world practice of FSI, as the values are within the range of frequencies not distorted by the band-pass filter system.

Table 2 presents the results of the LR analysis conducted on F2 of the five vowels. Here we can see better performance across vowels, with a minimum LR_{test} of 1.7 for /o/. /o/ is also the vowel with the lowest resolution of DS pairings, at 68.64%. Again, /a/ and /ɜ/ show best results for individual vowel LRs, with LR_{test} values of 5.2 and 4.1 respectively. DS pairings are, for F2 of all five vowels, better resolved than the SS pairings, with /a/ showing the minimum distance between the DS and SS rates of correct discrimination (81% SS, and 84% DS). /ɜ/ performs nearly as well as /a/, with a slightly lower LR_{test} value of 4.1. The reason for this is unclear, although it should be noted that F2 of /i/ and /ʉ/ perform approximately as well as /a/ and /ɜ/ performed using F1.

Table 2: LR results for F2

F2	/i/	/a/	/o/	/ʉ/	/ɜ/
SS	45.45%	81.82%	54.55%	63.64%	72.73%
DS	82.73%	84.55%	68.64%	77.27%	82.27%
LR_{test}	2.6	5.2	1.7	2.8	4.1

Table 3 presents the percentage of correct SS and DS resolutions using F3 of the five vowels. The only two vowels which show a DS resolution of over 80% are /i/ and /o/, which do not perform so well for F1 or F2. F3 of /ɜ/ does show a nearly 82% correct SS discrimination rate, equal to the highest SS discrimination looking at any other individual parameters.

Table 3: LR results for F3

F3	/i/	/a/	/o/	/ʉ/	/ɜ/
SS	36.36%	36.36%	54.55%	63.64%	81.82%
DS	80.45%	66.82%	80.45%	60.91%	67.73%
LR_{test}	1.9	1.1	2.8	1.6	2.5

A benefit of the LR approach is that it makes combining evidence easy (at least for uncorrelated variables). Since in FSI one is usually able to compare samples with respect to many variables, this is useful. In Table 4, the results of combining the LRs of different parameters together are shown (this is done by taking their product). This table shows a combination of all vowels F1, all vowels F2, all vowels F3 (so five parameters), then all vowels F1 and F2 combined (ten parameters), and finally all vowels for all formants combined together (fifteen parameters).

Table 4: Results for some combined LRs

	All F1	All F2	All F3	All F1&F2	All F1,F2 &F3
SS	36.36%	54.55%	36.36%	45.45%	36.36%
DS	96.36%	99.09%	94.09%	100 %	100%
LR_{test}	10	60	6.2	n.d	n.d

It can be observed in Table 4 that the possible magnitude of the LR_{test} score increases when parameters are combined. Using a combination of all vowels' F2 values results in a LR_{test} of 60, which is nearly twelve times as large as the highest LR_{test} using only a single parameter (5.2 for F2 of /a/). This is a useful result, as F2 is relatively unaffected by telephone bandpass (Rose, 2003). The lowest value given for a combination of five parameters is for All F3, with a value of 6.2. Note that this is still marginally larger than the highest LR_{test} for an individual parameter.

There is no LR_{test} value given for a combination of all F1 and F2 LRs, or for all fifteen parameters combined. This is because absolute discrimination of DS pairings was achieved using these two combinations. As the LR_{test} is a ratio of correctly discriminated SS pairings to incorrectly discriminated DS pairings, this means that we have a denominator of zero (i.e. no incorrectly discriminated DS pairings), resulting in an undefined LR_{test}.

This has some important implications for the application of FSI in Australian contexts using Bernard as a reference sample. First of all it shows that formants can be used to forensically discriminate same- from different-speaker pairs. Another important consideration is that while same-speaker comparisons for the above combinations are lower than for some of the individual parameters, this is balanced by the higher correct level of different-speaker comparisons. This has consequences in terms of the actual application of these methods to real FSI case work in Australia, or indeed anywhere using a similar legal framework for criminal prosecution. It can be seen to be more acceptable to provide evidence which may contribute to a guilty party being found not-guilty of an offence than to provide

Accepted after full paper review

evidence which (falsely) supports a prosecution same-speaker hypothesis.

At the same time, a method which cannot reliably generate an LR>1 for a SS pairing is not ideal – although as already mentioned that may be the wall effect in comparing identical samples. A balance, whereby a maximum number of correct SS pairings are resolved with a LR>1, while maintaining a high level of DS pairings resolved with a LR<1, should be sought, perhaps by seeing what kind of variance ratio gives an EER at threshold. Additionally, as the LRs are multiplied together, LRs making use of more parameters can have a theoretically larger magnitude LR. A secondary consideration in the combination is to maximise not only the DS and SS discrimination rate, but also the number of parameters in order to maximize the possible magnitude of the LR, and thus the strength of evidence supporting the hypothesis. This also provides some data to answer the secondary aim of the experiment, which is to examine the performance of the vowel F-patterns to see which are most useful and appropriate for actual FSI case work in Australia. Table 5 presents results of optimum combinations of individual parameters to meet these criteria.

Table 5: Optimum LR combinations using Bernard

	AllF2, /ɜ/&/a/F1	/ɜ/&/a/F1 /ɜ/&/a/F2, /ɜ/F3	AllF2, /ɜ/&/a/F1, /ʉ/&/ɜ/F3	AllF2, /ɜ/&/a/F1 /o/,/ʉ/& /ɜ/ F3
SS	63.64%	63.64%	63.64%	54.55%
DS	99.55%	99.55%	99.55%	99.55%
LR_{test}	140	140	140	120

Four combinations of parameters are presented in Table 5. The first three resolve SS and DS pairings with the same degree of accuracy – 63.64% of SS pairings (7 out of 11), and 99.55% DS pairings correctly discriminated (219 out of 220). The first column represents a combination of all five F2 parameters and the LRs calculated using F1 of /a/ and /ɜ/ (a total of seven parameters). The second column uses the LRs for all three formants of /ɜ/ in combination with the LRs for F1 and F2 of /a/. This is five parameters in all. The third uses the same parameters as column 1, but includes LRs for F3 of /ʉ/ and /ɜ/, for a total of nine parameters. The magnitude of the LR_{test} is here much larger than those given for any of the individual parameters, or those combinations listed in Table 4. This equates verbally to strong support for the LR>1 assuming SS hypothesis. The inclusion of the LRs for F3 of /o/ gives ten parameters; this is presented in the last column. Note the lower rate of SS discrimination, which lowers the LR_{test} (which still equates to strong support).

One method of graphically expressing both the magnitude of LRs calculated, and the relative performance of the SS and DS pairings is cumulative distribution functions (Drygajlo, Meuwly, and Alexander, 2003). Figure 1 presents the cumulative distribution functions of the Log_{10}(LR) values for the 5 (F1 & F2 in /a/, and all three formants in /ɜ/) and 9 (all vowels' F2, F1 in /ɜ/ & /a/, and F3 in /ʉ/ & /ɜ/) parameter combinations presented in Table 5. The 5 parameter combined LRs are shown in the left panel, with the 9 parameter LRs shown on the right. Each panel also shows the equal error rate for the test (labelled EER). Log_{10}(LR)s are shown along the x axis, with percentage of sample shown on the y axis.

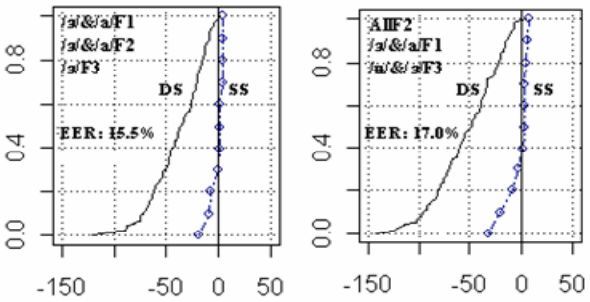

Figure 1: Cumulative Distribution Function of Log_{10}(LR) values for two optimal combinations of parameters.

In both panels the DS curve shows almost 100% of cases correctly discriminated at the threshold for the discrimination (0 in this case due to log transformation), with more than 75% of DS pairs being resolved with a Log_{10}(LR) < -25 for the nine parameter distribution function. SS pairings can be seen for both plots to have at least half of the curve on the positive side of the discrimination threshold. Note also the much larger magnitude of DS LRs than SS for both sets of LRs shown in Figure 1.

The maximum magnitude of SS Log_{10}(LR) is given by the nine parameter combination (column 3 of Table 5) – a value of 7, compared to -150 for the DS pairing with largest magnitude of DS Log_{10}(LR). This can be verbally expressed as very strong evidence to support the same-speaker hypothesis (this would be achieved with a Log_{10}(LR) greater than 3). The possible strength of evidence supporting the different-speaker hypothesis in the case of the DS pairings (with Log_{10}(LR)s of to a magnitude of -150) can be seen to be even stronger than this, although there is not another category of strength above very strong in the verbal scale used here (González-Rodríguez et. al, 2002).

Accepted after full paper review

4. Conclusion

The first aim of the experiment was to test the Bernard data set to evaluate its use in the actual application of FSI in Australia. The Bernard data has been shown to function well as a reference sample in Bayesian LR-based speaker discrimination and yields strong evidence using formants for FSI involving male speakers of AE. The collection of cultivated speakers will allow for further testing of accent type groupings of the Bernard data. The collection of a modern data set to compare results is another avenue for future investigation in this area; Cox (1999) provides data that may be useful for such a comparative test, though it is limited to speakers of General AE.

The LR_{test} values show that some combinations of the parameters can result in strong support for the hypothesis that the test can be used to successfully discriminate speaker pairs for male speakers of AE.

The second aim was to test the performance of the F-patterns for the first three formants of the /tense monophthongs/ of AE. Of the three formants, F2 was found to perform best in speaker discrimination, both for the individual vowels and for the combined LR given for a combination of all vowels for each formant, with F1 of /a/, and all formants of /ɜ/ also resolving speaker pairings relatively well. For almost all parameters or combinations of parameters, different-speaker pairings are resolved more successfully than same-speaker pairings.

For the three combinations of LRs found to give strongest support for the LR as a discriminant hypothesis, a combinations of 9 parameters - all five F2 LRs, in combination with F1 in /a/ and /ɜ/, and F3 in /ʉ/&/ɜ/ - was found to provide the strongest strength of evidence to support a hypothesis of same-speaker production of the two samples being compared (i.e. high magnitude LRs).

5. Acknowledgments

The experiments conducted for this paper would not have been possible without the time (and voices) of the eleven speakers who allowed me to record their speech at various times during 2003. Thanks must also be extended to the Australian National Dictionary Centre, which has funded my participation in this conference, as well as providing support, and to Phil Rose, who has provided much helpful advice.

6. References

Aitken, C.G.G. (1995) *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley: Chichester.

Alderman, T. (2004). Refining the Likelihood Ratio Approach to Forensic Speaker Identification – Effects of Non-Normality in the Background Distribution as Modelled with the Bernard Data for Australian English. Unpublished First Class Honours Thesis, Australian National University.

Bernard, J.R.L. (1970) "Towards the acoustic specification of Australian English". *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikations-forschung*. 23. pp113-28.

Bernard, J.R.L. (1989) "Quantitative aspects of the sound of Australian English". Blair & Collins (Eds.) A*ustralian English: The Language of a New Society.* University of Queensland Press: St. Lucia. pp 187-204.

Black, B., Francisco, J., & Saffran-Brinks, C. (1994) "Science and the Law in the Wake of *Daubert*: A New Search for Scientific Knowledge". *Texas Law Review* 72 (4): 715-802.

Champod, C. & Meuwly, D. (2000). "The inference of identity in forensic speaker recognition." *Speech Communication* 31: 193-203.

Cox, F. & Palethorpe, S. (1998) **"**Regional variations in the vowels of female adolescents from Sydney" in Mannell & Robert-Ribes (eds.) *Proceedings of the 5th International Conference on Spoken Language Processing,* Volume 6. The Australian Speech Science and Technology Association Incorporated: Canberra. pp. 2359-2362.

Cox, F (1999). "Vowel Change in Australian English." *Phonetica* 56: 1-27.

Drygajlo, A., Meuwly, D. and Alexander, A. (2003) "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition". *Proceedings of 8th European Conference on Speech Communication and Technology.*

González-Rodríguez, Ortega-García & Sánchez-Bote (2002). "Forensic Identification Reporting using Automatic Biometric Systems". In D. Zhang (Ed), *Biometric Solutions for Authentication in an e-World.* Kluwer Academic Publishers.

Horvath, Barbara (1985). *Variation in Australian English : the sociolects of Sydney.* Cambridge University Press: Cambridge & New York.

Kunzel, H. (2001) "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies" *Forensic Linguistics* 8(1): 80- 99.

Lindley, D. V. (1977). "A problem in forensic science." *Biometrika* 64(2): 207-213.

Mitchell, A.G. & Delbridge, A. (1965). *The Pronunciation of English in Australia (Revised Edition).* Angus & Robertson: Sydney.

Robertson, B. & Vignaux, G.A. (1995) *Interpreting Evidence.* Wiley: Chichester.

Rose, P. J. (2002). *Forensic Speaker Identification.* Taylor & Francis: London.

Rose, P.J. (2003). *The Technical Comparison of Forensic Voice Samples.* Issue 99, *Expert Evidence*, (series eds Freckleton, I and Sydney, H.). Thomson Lawbook Company, Sydney, 2003.

Accepted after full paper review