Bernard's 18 – Vowel Inventory Size and Strength of Forensic Voice Comparison Evidence

Phil Rose

Linguistics, College of Arts & Social Sciences, The Australian National University

philip.rose@anu.edu.au

Abstract

Eighteen vowel phonemes from Bernard's (1967) dataset of 171 Broad General and Cultivated male speakers are used to investigate how strength of forensic voice comparison evidence varies with number of different vowel phonemes available to compare samples. Logistic-regressively fused multivariate LRs are calculated from F-pattern and duration for vowel phoneme tuples of increasing size, and the best performing set of each tuple selected on the basis of its likelihood ratio cost function Cllr. The best performance, with EER of 1.41% and Cllr of 0.06, is achieved with all 18 vowels, showing that different speakers can still occupy very similar positions in heavily multiparametric formant space.

Index Terms: Forensic voice comparison, vowel formants, multivariate likelihood ratio, logistic regression, log likelihood ratio cost.

1. Introduction

In forensic voice comparison (FVC) a recording of a speech sample from an unknown voice, usually that of an offender, is compared with that of a known voice, usually the suspect. The task of the forensic expert is to estimate the likelihood ratio (LR) for the evidence, which is simply the probabilities of the evidence under the competing prosecution and defence hypotheses. In other words, they have to estimate how much more likely the differences between the suspect and offender speech samples are, assuming they have come from the same speaker rather than different speakers [6].

Currently there are two different approaches to LR-based FVC – *automatic* and *traditional* – distinguished primarily by the type of features used. Traditional features are those more closely associated with linguistic units, like F-pattern and F0. In traditional FVC the expert first scours the suspect and offender samples for comparable linguistic units, like vowel phonemes. The acoustic properties of these units, e.g. formants, are then quantified and LRs estimated. LRs for the different linguistic units are then combined to obtain an overall LR. Since traditional features can be expected to be correlated to a certain extent, for example, F2 and F3 in [i] vowels, simply combining the separate LRs in Naïve Bayes fashion can over- or underestimate the strength of evidence, and the combination of individual features must be done in such a way as to take the correlation into account. Currently this is done by using a multivariate LR to handle within-unit correlation (e.g. F2 & F3 in /i:/), and logistic regression to handle between-unit correlation (e.g. F3 in /ə:/ and /ʉ:/) [8].

Since the amount of correlation between traditional features turns out to be not particularly great, the more individual LRs one can combine, the greater the overall strength of evidence tends to be. Since LRs are usually estimated from vocalic F-pattern features, it would seem that the more contrasting vowels a language has, the better. The expected number of vowel phonemes for a Language is five [7]; AE has more than three times that. Moreover, again a typologically unusual phenomenon, seven of them are phonetically and/or phonologically diphthongal and thus contain more potentially individualising information [6].

The large number of vowel phonemes in Australian English is therefore potentially a 'good thing' for forensic voice comparison: it means the promise of getting good strengths of evidence. This paper investigates strength of FVC evidence as a function of vowel inventory size: it examines how the strength of evidence varies with the number and nature of vowels examined. The obvious theoretical question it asks is whether the large number of phonemes in AE is sufficient to correctly discriminate all same-speaker data in the Bernard corpus from different-speaker data. It also determines how the vowels rank in terms of strength of evidence, and what the useful subsets of AE vowels are for FVC. This last is an important practical question, because of one the things the investigator has to decide at the outset is whether forensic speech samples contain information that is likely to yield useful strengths of evidence for investigation or prosecution.

2. Data & variables

Painstakingly collected in the late sixties, Bernard's corpus contains *inter alia* information on the F-pattern (F1-F3) and duration of the eighteen (or nineteen if you include the marginal /ʊə/) vowel phonemes of 171 male Australian speakers, assigned to one of the three earlier conventional Broad, General or Cultivated accent categories of AE. This paper examines the discriminant performance of the five tense vowel phonemes (transcribed /iː ʉː əː aː oː/), the six lax vowels /ɪ e æ a o ʊ/, and the seven diphthongs /aɪ aʊ ɛɪ ɔɪ ɐʉ ɪə ɛə/. "Why use Bernard and not more up-to-date descriptions of Australian vowel acoustics?" – I am often asked – "Many Australian English vowels have changed since 1967". The answer is simple: Bernard's data are far from ideal, but there are still no available current descriptions of Australian vowel acoustics that match it for testing hypotheses in forensic voice comparison. Currently available summary descriptive-phonetic statistics based on normality assumptions from contemporaneous sampling cannot be used for forensic testing of the type illustrated here. The currency of Bernard's data for the research questions posed in this paper is also irrelevant. Hopefully this situation will change with the big ASC, and with researchers willing to share their data… .

Bernard recorded his subjects saying their vowels in /h_d/ words: once with the word in isolation, once with the word in stressed sentence-final position, and once prolonged. Since many speakers were unable to produced prolonged tokens, only the isolated and sentence frame tokens are used in this paper. There are therefore only two replicates per vowel per speaker. One replicate is treated as the suspect token; the other as the offender token, and the idea is to see how well a LR-

ISBN 978-0-9581946-3-1

based approach can discriminate between a pair of suspect and offender samples from the same speaker (i.e. target trials), and suspect and offender samples from different speakers (non-target trials). The presence of only one replicate per sample makes this a very tough task for discrimination, but this is mitigated by the fact that the samples are contemporaneous and are in highly comparable environments.

With 171 speakers, a maximum of 171 target trials and (14,535 * 4 partitions) = 58,140 different-speaker comparisons are possible. Only data from the first of the four partitions were used, yielding in all 14,535 non-target trials.

Conditions are never optimum in real FSR case-work, and it is important to incorporate if possible some reality constraints when experimenting with forensic discrimination. Thus discrimination was carried out under two main conditions called *optimum* and *quasi-realistic*, the former making use of as many variables as possible; the latter selecting only those variables that one might expect to be able to quantify under some real case-work conditions, and also taking into account the vowels' frequency of occurrence.

Bernard sampled F-pattern from spectrograms at onset, first target, second target if any, and offset. He also measured the duration of any onset perturbation, any target, any transition between first and second targets, and any offset perturbation. *Optimum* variables were selected from these measurements as follows. The seven phonological diphthongs /ɛɪ aɪ aʊ ɔɪ ʁʉ ɪə ɛə / were quantified with all three formants at both targets, as well as duration of the first target and duration between first and second target: a total of 8 variables per diphthong. /iː/ and /ʉː/ usually have phonetically diphthongal allophones in AE and were quantified in the same way as the phonological diphthongs, except that duration was not possible with /ʉː/, where addition of duration data resulted in matrix singularity. The monophthongs /əː aː oː ɪ e æ a o/ were quantified with all three formants and nuclear duration. Not surprisingly, many speakers lacked F3 for /ʊ/, and it was quantified just with F1 and F2. Thus the optimum model was based on 104 parameters. Since most FSR samples are from telephone speech where the F1 of high and mid vowels is compromised, the *quasi-realistic* condition omitted all F1 values at high or upper-mid targets. So for example in /aɪ ɛɪ ʁʉ ɔɪ aʊ iː ʉː/, F1 on the second target was omitted. F3 in /o/ and /ʊ/ is usually weak and difficult to measure, so it too was omitted (many of Bernard's speakers indeed lacked F3 for /ʊ/, although most had intact F3 for /o/). F3 was otherwise retained, since even for high front vowels it lies typically somewhat below a nominal upper bandpass of 3.5 kHz. Duration data were also discarded: they are unlikely to be of use in any but the most tightly controlled conditions.

3. Processing

3.1.1. Multivariate Likelihood Ratio

Kernel density multivariate LRs were estimated separately for each of the 18 vowel phonemes using the formula developed at the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* [1]. This formula is particularly suited for the sparse data in this paper since it estimates within-speaker variance over both replicates. For each vowel a set of target LRs and a set of non-target LRs were generated, representing the 171 same-speaker comparisons and 14,535 different speaker comparisons respectively.

3.1.2. Calibration

The MVLRs were then calibrated using Brümmer's *Focal* tool-kit code [4]. Calibration is an optimization which monotonically transforms the LR values on the basis of the actual hypothesis they represent [3]. To illustrate its effect, figure 1 shows, with a Tippett plot, the results of the MVLR analysis on /ɛə/, before and after calibration. In figure 1, the two curves increasing to the right represent the cumulative distribution of $\log_{10}$LRs from different-speaker comparisons. The two curves increasing towards the left are the same-speaker $\log_{10}$LRs. It can be seen that the EER is about 27%, so that about 73% of the different-speaker LRs lie below the threshold of 0, and about 63% of the same-speaker LRs lie above. The "scissoring" effect of the calibration can be clearly seen. Its most important result is the reduction in the magnitude of the "incorrect" LRs. With uncalibrated values it was possible to find a same-speaker comparison evaluated with $\log_{10}$LRs up to just below -4, and a different-speaker comparison evaluated with a $\log_{10}$LR up to about 2. These are reduced by up to two orders of magnitude after calibration: for obvious reasons, this is seen as a desirable result. Calibration has also reduced the greatest strength of evidence achievable. For example, 20% of uncalibrated different-speaker $\log_{10}$LRs have values less than ca. -2; this is reduced to ca. -1 for calibrated LRs.

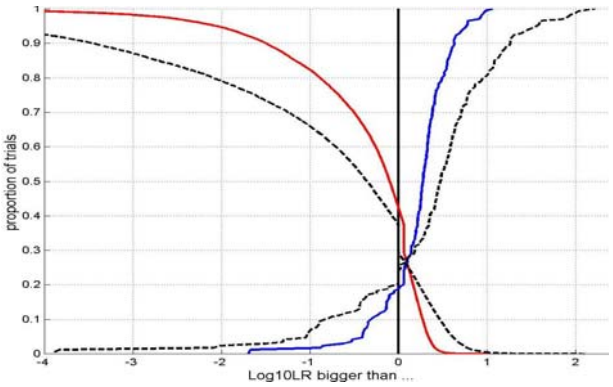

Figure 1: *Reliability (Tippett) plot for optimum /ɛə/ log_{10}LR distributions, showing effect of calibration. Dotted lines = uncalibrated LRs, solid lines = calibrated LRs.*

3.1.3. Log Likelihood Ratio Cost

$$C_{llr} = \frac{1}{2}\left(\left[\frac{1}{N_{Hp}}\sum_i \log_2\left(1+\frac{1}{LR_i}\right)\right]+\left[\frac{1}{N_{Hd}}\sum_j \log_2\left(1+LR_j\right)\right]\right) \quad (1)$$

The performance of LR-based detection systems like the one in figure 1 is currently evaluated with the Log Likelihood Ratio Cost (Cllr) [3], and this metric was also used in this paper. The formula for this simple scalar metric is given at (1), where it can be seen to consist of the mean of two hypothesis-dependent logarithmic functions. The left-hand term evaluates the performance of all same-speaker LRs; the right-hand one the performance of all different-speaker LRs. The purpose of Cllr is to severely penalize highly misleading LRs. For example, a different-speaker LR of 1000 (misleadingly and dangerously indicating that one would be 1000 times more likely to get the difference between the suspect and offender samples had they come from the same speaker) attracts a value of $\log2(1 + 1000) \cong 6.909$. Since Cllr effectively does not

reward correct cases, even if they involve high LRs, this value of 6.909 then has a high contribution to the mean of all the different-speaker LRs and the overall Cllr value. The uncalibrated Cllr value for the /ɛə/ data in figure 1 is 0.9119, which improves to 0.8157 for the calibrated data (Cllr values below unity are considered good, but neither of these is particularly good).

3.1.4. Fusion

One of the aims of this paper is to find out how discrimination performance is affected when LRs from different vowels are combined. Within-segment correlation, for example between formants, can be handled by the use of multivariate likelihood ratios, but between-segment correlations remain a problem. The current solution is to adopt methods from automatic speaker recognition and use logistic regression fusion [9]. Although, as will be seen, this works pragmatically to give a much better result in terms of Cllr and EER for combined than individual vowels, it is not without theoretical problems. This is because logistic regression fusion does not operate on correlation between variables, as MVLRs, but takes into account correlations between the ensuing LRs. It is therefore possible - although this has not yet been shown - that two segments which are not correlated by virtue of their internal structure and which therefore should be naively combined, nevertheless have LRs which do correlate. In this case, the logistically regressive combination will underestimate the combined strength of evidence.

4. Results – individual vowels

Table 1. *Results for individual vowels*

/V/	f_{RP} (%)	EER (%)		Cllr	
		opt.	*real.*	*opt.*	*real.*
ɛɪ	1.71	15.25	23	0.493	0.7388
aɪ	1.83	16.5	25.5	0.5345	0.7271
iː	1.65	17	21	0.5919	0.7642
aʊ	0.61	19.5	26.5	0.600	0.7815
ɔɪ	0.14	20	29.5	0.5956	0.7532
ɐɥ	1.51	21	25	0.6748	0.8041
ɪə	0.21	22.5	29.5	0.6569	0.8138
æ	1.45	22.5	25	0.7224	0.8328
ɛə	0.34	27	28	0.8157	0.9235
ɒ	1.37	27	35	0.8243	0.8859
oː	1.24	27.5	30	0.8442	0.8159
ɪ	8.33	28	32.5	0.7758	0.8725
a	1.75	28	33	0.8397	0.9023
aː	0.79	29	30	0.8173	0.8043
əː	0.52	29.5	30	0.8171	0.8602
e	2.97	30	33	0.8217	0.8694
ɯː	1.13	30	35	0.903	0.8814
ʊ	0.86	33	37	0.9256	0.8685

One of the questions posed in this paper was how individual vowels rank in terms of potential strength of evidence. Table 1 lists the EER and Cllr values of the LR-based analyses of the 18 separate vowels under both optimum and quasi-realistic conditions, ranked according to optimum EER. The percent textual frequency of the vowels' RP cognates (f_{RP}) from [5] is also given. There are no real surprises: the optimum conditions yield better results than the realistic, and the vowels with the highest individual identifying potential are the closing diphthongs /ɛɪ aɪ iː aʊ ɔɪ ɐɥ / with EERs between ca. 15% and 20%. Of these it is encouraging to see that they are also among the most common of the vowels after /i/ and /e/. Of note is the lax vowel /æ/ which, with an EER of 22.5%, is close to the diphthongs, despite the fact that it has much less information, with only one set of "target" formant measurements. The remaining vowels have EERs between ca 27% and 33%. These results show that if the investigator has a choice, closing diphthongs, /iː/ and /ae/ are likely to be the best vowels to quantify.

5. Results – combined vowels

Table 2. *vowel tuple composition ranked by Cllr.*

tuple	composition	tuple	composition
1	**/eɪ/**	10	9 + **/ɪ/**
2	**/aɪ oɪ/**	11	10 + **/æ/**
3	2 + **/iː/**	12	11 + **/ʊ/**
4	3 + **/oʊ/**	13	12 + **/oː/**
5	4 + **/ɪə/**	14	13 + **/a/**
6	5 + **/eɪ/**	15	14 + **/o/**
7	6 - **/eɪ/** + **/ɯː/** + **/aʊ/**	16	15 + **/ɛə/**
8	7 + **/eɪ/**	17	16 + **/e/**
9	8 + **/aː/**	18	17 + **/əː/**

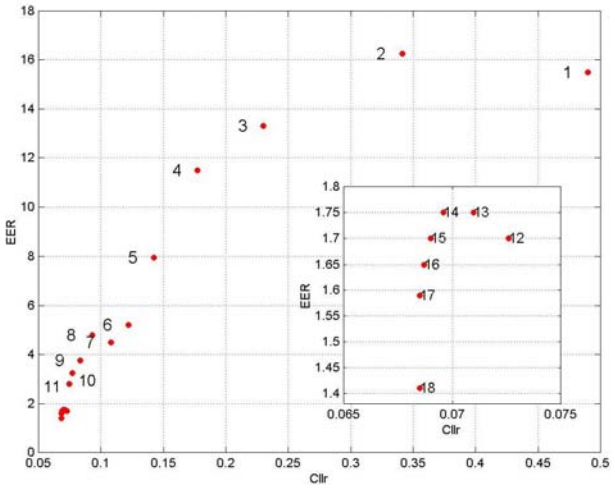

Figure 2: *EER plotted against Cllr for all 18 optimum LR vowel tuples. Numbers indicate vowel tuple. Insert shows details at bottom left corner.*

To investigate the second question posed by this paper - what happens to the strength of evidence when vowels are combined - the individual vowels' LRs were combined by progressively increasing the size of vowel tuples from two to the full 18, determining which set of vowels within each tuple gave the best Cllr. The results of this experiment are given in table 2. Table 2 shows for example that the best two-vowel combination was /aɪ/ and /ɔɪ/, and the best triplet was these two vowels plus /iː/. Generally, but not always, the $n + 1$ tuple contained the n tuple as a subset. As an example, as already indicated, the best single vowel according to Cllr was /ɛɪ/ but the best dublet did not contain /ɛɪ/. This is presumably because of high correlation between the LRs of /ɛɪ/ and one or both of /aɪ/ and /ɔɪ/. Another example is /əː/, which is the last

vowel to participate, again presumably because its LRs have the greatest correlation with all the other vowels. The performance of the 18 combinations in terms of EER and Cllr is shown in figure 2. It can be seen that both EER and Cllr improve with LRs from an increasing number of vowels, from ca. 16%/0.35 for the best two vowels, through ca. 8%/0.15 for five vowels, to ca. 1.4%/0.068 for all 18. Diminishing returns obtain, however: you do not get much improvement in EER or Cllr after about 12 vowels. Generally, the EER is proportional to the log of the Cllr.

The third question of this paper was whether, using combined LRs from all 18 vowels, it was possible to completely discriminate all same-speaker data from different-speaker data. We already know from figure 2 that complete separation is not achieved, although the EER is about 1.4% and the Cllr is also quite low, at 0.06. The details of the performance with all 18 vowels is shown in the Tippett plot in the top panel of figure 3. The same-speaker pairs are resolved very well, with only one speaker out of 171 having a difference between their two replicates that was more likely to be a different-speaker difference (but not much more – speaker 102's $\log_{10}$LR for all 18 vowels was only -0.2, and not a worrying error).

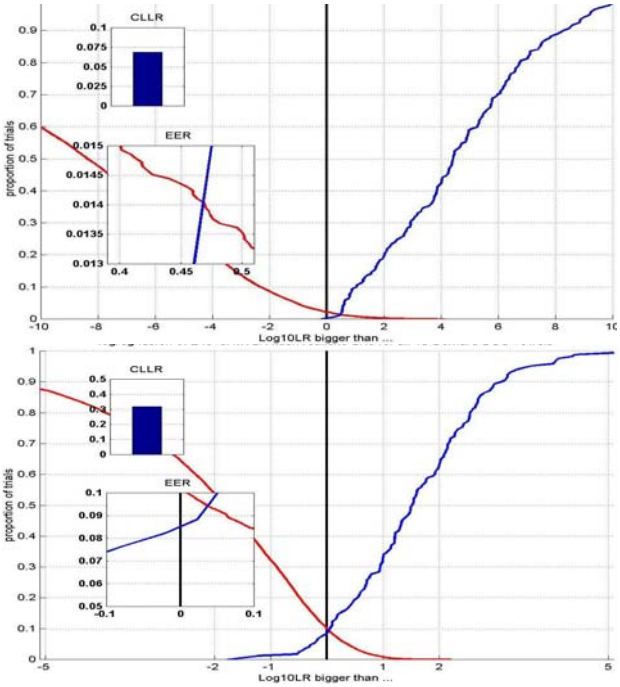

Figure 3: *Tippett plots for combination of LRs of all 18 vowels. Top = optimum LRs, bottom = quasi-realistic LRs. EER insert shows detail near intersection of curves.*

It is a somewhat different story for the different-speaker LRs. Although the error rate is quite low at 2%, the spread of different-speaker $\log_{10}$LRs that were separated by differences that were more likely had they come from the same-speaker is still rather large. About a quarter of a percent of different-speaker pairs had "incorrect" $\log_{10}$LRs of between 2 and 3: i.e. between one hundred and one thousand times more likely assuming same-speaker provenance. About 0.06 percent had $\log_{10}$LRs of between 3 and 3.5; four different-speaker pairs (ca. 0.03%) had $\log_{10}$LRs between 3.5 and 4.0; and one different-speaker pair had a $\log_{10}$LR of 4.1. This is about *12.5 thousand times* more likely assuming same-speaker provenance. When one considers that this value is based on 104 parameters from 18 separate vowels, it is salutary to be reminded that there were 2 speakers in the 171 speaker dataset (75 & 116) that had values for all their vowel parameters that were sufficiently similar and atypical to be evaluated as a same-speaker pair with extremely high confidence.

To a certain extent, of course, one cannot get a good idea of the forensic potential of vowel acoustics by looking at the optimum scenario, as above. It does not matter how good a vowel is, or a group of vowels, if you are unlikely to meet them often in forensic speech samples. In a second way of combining the vowels, therefore, their quasi-realistic LRs were used, with vowels grouped according to their textual frequency (this was given in table 1). Thus the most common vowel pair is /ɪ e/, the most common triplet is /ɪ e aɪ/ etc. The bottom panel of figure 3 shows their Tippett plot. The EER for the realistic discrimination is 9.4%, and Cllr 0.32, and so as expected it is not as good as with the optimum LRs. However, as can be seen, the extent of "bad" different-speaker comparisons is much reduced: with this model, at least speaker 75 is no longer likely to get convicted of the crime perpetrated by speaker 116.

To return, finally, to the typological question posed at the beginning of this paper: yes, it probably *is* useful having offenders speaking Australian English, (or probably any Germanic language for that matter) with its large number of vowel phonemes. However, offenders speaking languages with smaller vowel inventories should not relax: such languages might be expected to have a larger number of vowel replicates per phoneme for LR estimation, which is also a good thing.

6. Acknowledgements

This paper was written as part of *Australian Research Council Discovery Grant* No. DP0774115. I have followed my reviewers' critique to make, I hope, a better paper: many thanks! My biggest thanks must go to Bernard.

7. References

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

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

[3] Brümmer, N. & du Preez, J. "Application independent evaluation of speaker detection", Computer Speech and Language, 20(2-3), 230-275, 2006.

[4] Brümmer, N. "Focal Toolkit" http://www.dsp.sun.ac.za/nbrummer/focal

[5] Gimson, D. An Introduction to the Pronunciation of English, London: Edward Arnold, 1962.

[6] Gonzalez-Rodriguez J., Rose P., Ramos, D., Torre, D. & Ortega-García, J. "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", IEEE Transactions on Audio Speech and Language Processing 15(7), 2104 – 2115, 2007.

[7] Lass, R. Phonology: an Introduction to Basic Concepts, Cambridge: CUP, 1984.

[8] Rose, P. "The Effect of Correlation on Strength of Evidence Estimates in Forensic Voice Comparison: Uni- and Multivariate Likelihood Ratio-based Discrimination with Australian English Vowel Acoustics", International Journal of Biometrics 2/14: 316-329, 2010.

[9] Pigeon, Stéphanie, Druyts, Pascal, & Verlinde, Patrick, "Applying Logistic Regression to the Fusion of the NIST '99 1-Speaker Submissions", Digital Signal Processing 10/1-3, 237-248, 2000.