The Intrinsic Forensic Discriminatory Power of Diphthongs

Phil Rose

Phonetics Laboratory, Linguistics (Arts), Australian National University
philip.rose@anu.edu.au

Abstract

This paper describes an experiment investigating how well same-speaker speech samples can be discriminated from different-speaker speech samples using acoustic parameters from Australian English diphthongs. A two-level kernel density multivariate likelihood ratio is used as a discriminant function on five of the diphthongs of the 171 speakers of the Bernard data base: /aɪ ɛɪ ʊʉ ɪə ɛə/. Comparing samples with just one token per diphthong each, an EER of ca.10% is obtained. It is concluded that diphthongal F-pattern can yield useful strength of forensic speaker identification evidence.

1. Introduction

Over about the last twenty years much attention has been given to the proper evaluation of forensic evidence, resulting in a major paradigm shift in many areas of forensic identification science, including forensic speaker recognition (Rose 2006a). The paradigm shift was ultimately due to the post-1968 "new evidence scholarship" debate, and the increased incidence of statistical evidence associated with forensic DNA profiling. As a result, it is now well known that however much the court or the police may desire otherwise, the forensic identification expert cannot logically or legally quote the probability of the hypothesis given the forensic evidence (Rose 2006a). Instead, they must estimate the strength of the evidence by calculating its likelihood ratio (LR): the ratio of the conditional probabilities of the evidence under competing prosecution and defence hypotheses. In forensic speaker recognition, therefore, the expert has to estimate how much more likely one is to observe the differences between the offender and suspect speech samples assuming that they have been spoken by the same speaker than by different speakers.

Since the LR quantifies how much more likely the evidence is under the prosecution than the alternative, defence, hypothesis, it can be used as a discriminant function to test the discriminatory potential of different kinds of forensically common evidence. If the evidence is of any discriminatory value, the LR should be greater than unity for same-subject data, and less than unity for different-subject data. At the same time, such testing provides a demonstration of the viability of the method. This is absolutely crucial forensically in the wake of the *Daubert* ruling on admissibility of scientific evidence: that approaches should have been tested. It is also consistent with the injunction that forensic identification science should emulate the DNA profiling approach, which has long used LRs.

Quite a lot of this kind of testing has already been carried out in both Forensic Automatic Speaker Recognition, and Traditional Forensic Speaker Recognition. The forensic discriminatory power of formants in the long monophthongs of Australian English has been quite well researched over the last few years. Alderman (2005) for example showed that non-contemporaneous same-speaker speech samples could be discriminated from different-speaker samples using the formants of the five AE long monophthongal phonemes /i a ʉ ə o/. Rose (2006b) showed that discrimination can also occur when the correlation between the individual variables is taken into account. Loakes (2006), examining the performance of both long and short monophthongs in the natural speech of Victorian identical and non-identical twins, found that different-speaker pairs can be relatively easily discriminated; but that some twin pairs, not surprisingly, were not easily discriminable on the basis of mean F-pattern alone (although they could be very well discriminated with other parameters).

Up to now, likelihood ratio-based forensic discrimination experiments have concentrated on monophthongs. Although some speaker-recognition experiments have been done using diphthongal formants (McDougall 2004, 2006), or formant trajectories (Ingram et al. 1996), they have not used approaches which are interpretable forensically. Thus the forensic discriminablity of diphthongs remains unexplored. It is the aim of this paper to make a start.

It is reasonable to assume that diphthongs contain more potentially useful individualising information than monophthongs. Spectrally, two targets are involved, each with up to three formants. Whether this constitutes double the information of monophthongs depends on how much correlation exists between the formants of the two targets. Speakers can also differ in the nature of the transition between the two targets (McDougall 2004, 2006). In addition with diphthongs there are duration parameters that can be explored, for example, the relative timing of the targets. The aim of this paper, then, is to investigate aspects of discrimination with diphthongs. In real-world forensic case-work, it is obviously useful to know which parameters, and which diphthongs, are likely to yield the strongest evidence. This paper asks, therefore: (1) to what extent diphthongs can be used to discriminate same-speaker from different-speaker speech samples; and (2) which parameters of which diphthongs are the best.

2. Data

Australian forensic speech scientists are lucky in having, if not an ideal, at least a substantial resource for testing – the

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Bernard corpus (Bernard 1967). Collected in the late sixties, this contains information on the F-pattern (F1-F3) of not only all eleven monophthongal vowel phonemes but also the seven diphthongal phonemes of about 170 male Australian speakers, assigned to one of the three earlier conventional Broad, General or Cultivated accent categories of AE. This paper examines three falling and two centering diphthongs: /aɪ/ as in *by*, /ɛɪ/ *bay*, /ɐʉ/ *bow*, /ɪə/ *beer* & /ɛə/ *bear*.

Bernard recorded his subjects saying their diphthongs in /h_d/ words: once with the word in isolation, and once with the word in stressed sentence-final position. So there are only two tokens per diphthong per speaker. He sampled their F-pattern from spectrograms at four places in the Rhyme: onset, first target, second target, and offset. Bernard was also far-sighted enough to include durational information. Thus for each diphthong he measured the duration of any onset perturbation; the duration of its first target if any; the duration of the transition between the first and second targets, and the duration of any offset perturbation. This information permits the plotting of a crude time course of the diphthongal F-pattern. Figure 1 shows such a reconstructed plot, for the F-pattern time course, plotted as a function of raw duration, for /aɪ/. (The mean of each speaker's two tokens is plotted, apart from a couple of speakers for whom only one token is recorded.) The overall F-pattern can be seen to be typical of a diphthong with a fairly low back first target (as shown from the relatively high F1 and low F2), and a high front second target (reflected in the low F1 and high F2). The relatively high F3 reflects lack of lip rounding. The prolonged values of the first target, which last for about half the duration of the Rhyme, are indicative of a falling diphthong, i.e. one with prominence on the first target. A considerable amount of between-speaker variation can be seen in all aspects of the F-pattern time-course, with the first target of F3, and the second target of F2 and F3 varying the most.

In addition to acoustic measurements, each diphthongal token was also classified according to several auditory features, including whether it, as opposed to its speaker, sounded Broad General or Cultivated. In this paper, there is no special treatment as to accent, tokens from all speakers being pooled. It is of course possible to investigate whether, say, diphthongs classified as Broad perform better than General; or whether the diphthongs spoken by speakers classified as Broad perform better than those spoken by General speakers.

3. Experimental Variables

Bernard's quantification contains many potential parameters from which to chose. There are (4 sampling points * 3 formants =) 12 separate acoustic values for each token, as well as five separate duration measurements (six if one counts overall Rhyme duration derivable from their sum). Various combinations of these parameters are of interest for discrimination purposes. In likelihood ratio-based traditional forensic speaker recognition, vowel formant centre frequencies have been shown to yield moderately strong evidence, and so it is of obvious interest to see how well speakers can be discriminated just with their diphthongal

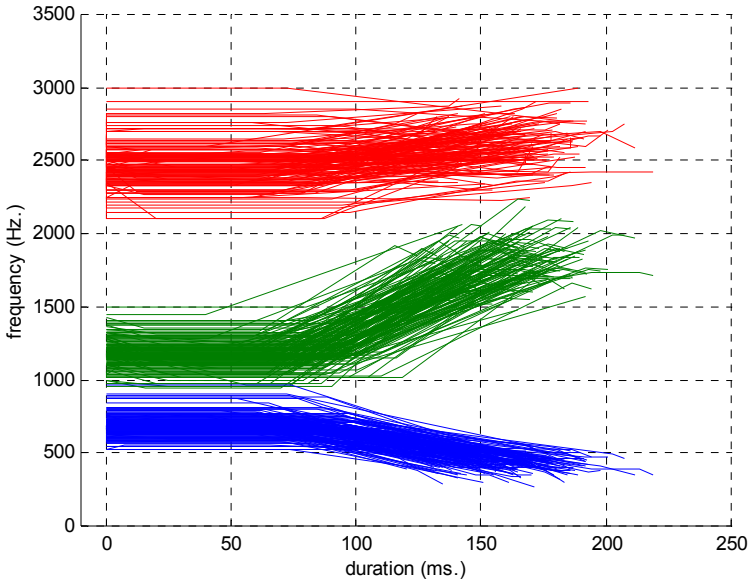

Figure 1 : Time course of /aɪ/ F-pattern in 169 of Bernard's male Australian speakers.

formants. Of these, it is the F-pattern values at the two diphthongal targets that are of primary interest. In most cases there is very little difference between the F-pattern values at onset and the first target on the one hand, and offset and the second target on the other. It is thus reasonable to use just the six F-pattern values at the two targets as spectral parameters, and ignore the onset and offset values. This – F1 to F3 at both targets – therefore constitutes one experimental condition, which will be called the *all formants* condition. Since most FSR samples are from telephone speech where the F1 of high vowels is compromised, a second, more realistic, experimental condition is to exclude all potentially compromised F1 values. For /aɪ/ this will be the F1 of the second target. This, with one less observation, will be called the *realistic formants* condition.

As already mentioned, Bernard's diphthongal data contains a wealth of durational information, and it is of interest to see how much this contributes to discrimination. Figure 1 suggests a certain amount of between-speaker variation in the duration of the first target and the duration of the transition between the first and second targets, and these were chosen as two additional duration parameters. This will be called the *raw duration* condition. Since it is unlikely that absolute duration measurements will be of use in real-world casework, in addition discrimination was attempted using duration values normalised by dividing by the Rhyme duration. This will be called the *normalised duration* condition.

Of interest also is of course how well the combined spectral and duration parameters perform. Two further combined conditions were thus tested: (1) *all formants* combined with *raw duration*), called *all combined* condition; and (2) *realistic formants* combined with *normalised duration*, called *realistic combined* condition. Discrimination was thus carried out under the above six experimental conditions.

4. Processing

Diphthongs (and of course voices in general) are heavily multidimensional: that is why they are of such potential

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

interest in FSR. One downside of using diphthongs forensically is that, although they have many potentially useful parameters, not all of them will be uncorrelated. The /aɪ/ diphthongal data, for example, shows significant positive correlations between several formants: F2 and F3 at the second target (r = 0.378); between F3 at the first and second targets (r = 0.354); between F1 at the first target and F2 at the second target (r = 0.291), and between F1 and F2 at the first target (r = 0.567). This means that an overall likelihood ratio for a comparison using more than one diphthongal parameter must be able to take correlation into account, otherwise the strength of the evidence will be grossly overestimated (Rose 2006b, Rose Lucy & Osanai 2004). The formula for the LR used in this paper was derived at the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* precisely as a solution to the non-trivial problem of estimating the strength of evidence when predictor variables may be correlated (Aitken & Lucy 2004).

The LR formula treats the variables for which a LR has to be estimated – for example a set of measurements of frequencies for different vowel formants – as multivariate data, 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. It was decided to extract kernel density LRs, since it is known that some distributions can deviate significantly from normality.

It is very important for this paper to note that the MVLR formula at present only accommodates two levels of variance: between- and within-subjects, and is thus somewhat unrealistic for FSR, where at least a third level of variance – between non-contemporaneous sessions – must usually be assumed. This is because the within-speaker variance between sessions is usually greater than within a session, and speech samples to be compared forensically are usually separated by more than a few seconds or minutes. (Usually, but not invariably: some phone calls are timed to be separated by a matter of a few seconds, and clearly would constitute 'same-session' data.)

The approach is probably also unrealistic for speech in requiring the within-subject variance to be modeled normally.

The formula for the multivariate kernel density LR is reproduced from Aitken & Lucy (2004: 116, 117) at (1) (its numerator) and (2) (its 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.

The general approach in this paper is the same as that used in Aitken & Lucy (2004) for testing trace evidence (elemental ratios of glass fragments), and can be explained once again taking /aɪ/ as an example. Each trial involves the comparison of one set of values from a speaker – all three formant values at both targets, say – with either the other set of values from the same speaker or another set of values from a different speaker in the corpus. The difference Δ between both sets of values under comparison is evaluated against a model of the distribution of same-speaker data in the corpus to determine the probability of getting Δ assuming the two values were spoken by the same speaker – the numerator of the LR $p(\Delta|H)$; and against a model of the distribution of all speakers in the corpus to determine the probability of getting Δ assuming the two sets were spoken by different speakers – the LR denominator $p(\Delta|\sim H)$. This kind of LR-discrimination experiment, where the test data also constitute the reference population, can be called *intrinsic*, whence the paper's title.

Formant and duration data from /aɪ/ were available from 166 speakers. This meant that 166 same-speaker comparisons, or target trials, and 13,695 different-speaker, or non-target trials, could be made. A very important aspect of this comparison, and one that differs from Aitken & Lucy (2004), is that it involves only two tokens of /aɪ/ per speaker, since that is all the Bernard data contains. Thus for same-speaker

numerator of MVLR = (1)

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

denominator of MVLR = (2)

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

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 = $\left(4/\left(2p+1\right)\right)^{1/(p+4)}m^{-1/\left(p+4\right)}$
$\bar{y}_l = \bar{y}_1, \bar{y}_2$ = 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.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

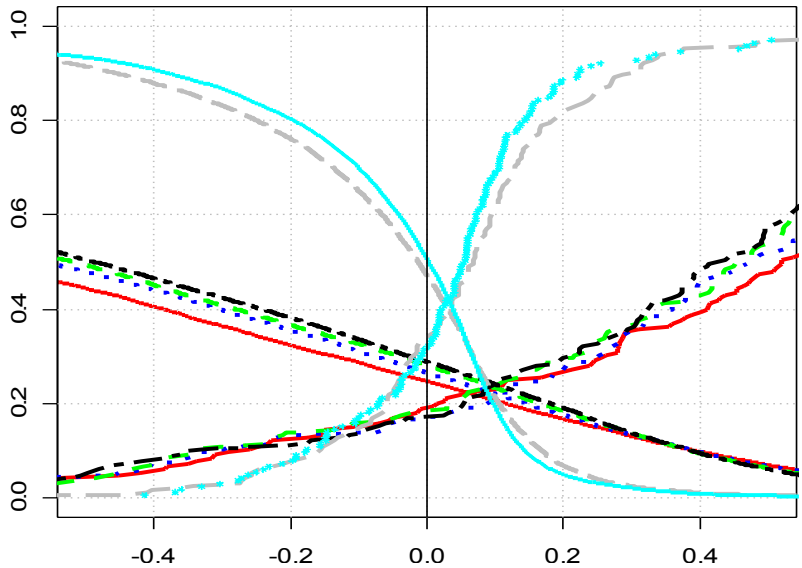

Figure 2: Tippett plots for results of MVLR discrimination with /aɪ/. Horizontal axis is $Log_{10}LR$, vertical axis is probability. Red = all comb.; blue dots = realistic comb.; green dashes = all formants; black dot-dash = realistic formants. Grey dashes = raw duration; cyan = normalised duration.

comparisons one /aɪ/ token is treated as the suspect data, and the other as the offender. This of course makes the discrimination very difficult from the outset: usually one works with considerably more than one replicate per sample, since statistically that increases the representativeness of samples, which helps the discrimination.

5. Results for /aɪ/

Forensic Speaker Discrimination results are conventionally presented in so-called Tippett, or reliability, plots, which 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 equal to or bigger than a given abscissa LR value. This enables a clear statement of the probability of error – another important *Daubert* criterion. So a typical statement for the court might be 'these graphs show that if I evaluate the evidence with this approach, and get a LR of 100, I shall be wrong 5% of the time if I take this LR to support the defence hypothesis that the samples come from the same speaker'.

Normally in Tippett plots, the same-speaker LRs are plotted inversely, but this makes it difficult to see important details like the equal error rate (EER), and is avoided here. Also, LRs for different-speaker comparisons can get very big if, as here, speakers from all parts of the distribution are compared. 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 ($Log_{10}LR = 0$).

Figure 2 shows the results of the MVLR discrimination with /aɪ/, using the six different experimental variables listed above. Note that only a small portion of the results, between $Log_{10}LR$ = -0.5 and 0.5, is shown. The different-speaker, or non-target plots decrease towards the right, and the same-speaker, or target plots decrease towards the left. It can be seen that there are two groups of lines. Those representing the duration-only discriminations are in cyan and grey and form an X-shape in the middle of the graph with an EER, just to the right of the $Log_{10}LR = 0$ threshold, of between 35% and 45%. Those representing discrimination with formants only, or combinations of formants and duration, form a flatter X with an EER between 20% and 30% located slightly further to the right of threshold. Numerical results of the discrimination – the EER and its location on the x axis (loc.) – are given in table 1.

5.1. Calibration

An important feature of the results, also easily seen in figure 2, is that all curves have EERs effectively the same as threshold ($Log_{10}LR = 0$). The furthest EER away from threshold is 0.109, which corresponds to a LR of 1.286, effectively no different from LR = 1. This result is worth commenting on in the light of recent research into the evaluation of speaker recognition systems with cost-based metrics using strictly proper scoring rules derived from forecasting (Brümmer & du Preez 2006). One of the problems which LR-based forensic automatic speaker recognition systems have to confront is poor calibration: they may show very good discrimination, but their EERs are often located a long way from the threshold. It is interesting to see, therefore, that calibration does not seem to be a problem with LR-based discrimination using *traditional* features, as shown here, or in Rose (2006b). This may be related to the fact that traditional LR-based discrimination makes use of a analytically derived LR formula, like that at (1) and (2), whereas automatic approaches use an empirical number-crunching approach. Calibration may cease to be a problem if an analytically derived LR formula is used, therefore. At any rate, it is not a cause for worry in this set of experiments. This allows us to focus on the discriminatory aspect of the performance.

5.2. Discrimination

The results show firstly that combination of the two duration parameters (the duration of the first target, and the duration of the T1-T2 transition) is a poor discriminator on its own, whether normalised or not. There is little difference between the raw and

Table 1: Results of discrimination using /aɪ/

	all combined	realistic combined	all formants	realistic formants	raw duration	normalised duration
EER	22.29%	21.66%	23.52%	24.16%	39.66%	42.08%
loc.	0.061	0.109	0.0965	0.0981	0.0327	0.0316

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

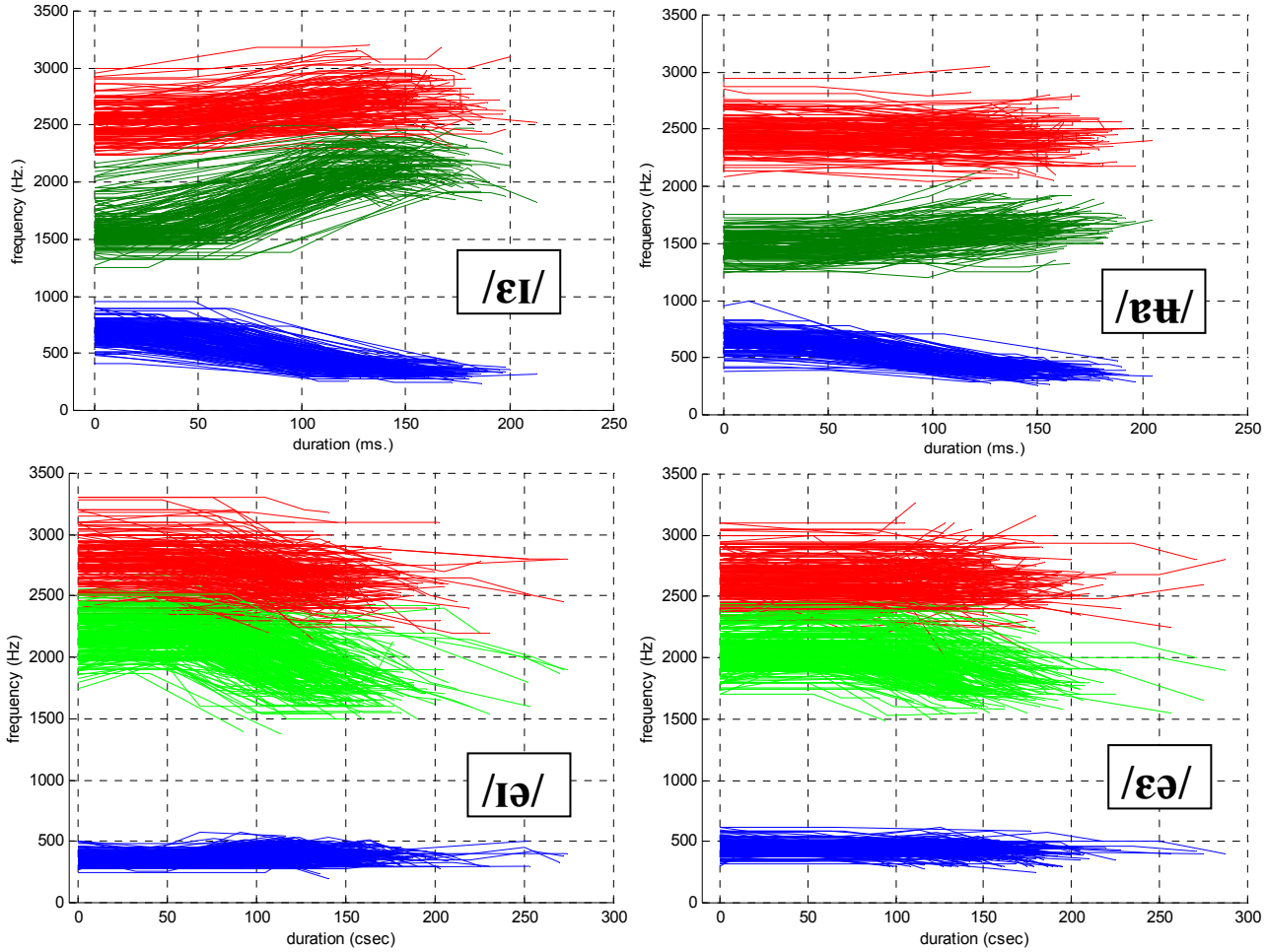

Figure 3: Multispeaker F-pattern plots for four other Bernard diphthongs tested.

normalised parameters, and with EERs of about 40% each, a LR at about threshold would be wrong about 40% of the time. Neither is any great strength of evidence to be expected from duration alone: at the best it will be between 3 and 5 times more likely under one hypothesis than the other. Nevertheless, this should not obscure the fact that there is a certain amount of individual discriminatory power in these durational parameters. Figure 2 shows for example that, given the fairly steep slope of the Tippett plots for duration, there are very few different-speaker pairs that are evaluated with a LR > $\text{Log}_{10}0.5$, and therefore a Log_{10} of about 0.5 from duration alone would involve a very small probability of error.

Not surprisingly, results are improved considerably when formants are incorporated, although there is really very little difference between the results for the different experimental conditions. Making use of formants and duration information – i.e. the combined conditions – gives EERs of 22%; just using formants on their own increases the EER to 24%. So once again one sees that duration makes a little contribution to discriminablity. The best EER, but only just, is obtained with the *realistic combined* condition, i.e. excluding F1 for the offglide, and using normalised duration.

6. Results with other diphthongs

Forensic discrimination tests were run using the MVLR formula on four other diphthongs from the Bernard set: /ɛɪ oʊ, ɪə ɛə/. Their raw values are plotted in figure 3, which shows lots of between-speaker variation in F2 and F3, but much less in F1. This suggests that perhaps F1 should simply be ignored in future experiments. Results are given in table 2, which shows the number and percent correct of same-speaker and different-speaker trials for all five diphthongs separately, and for all five combined. The combined result was obtained in Independence Bayes fashion, by summing the individual diphthong Log_{10}MVLRs. As usual, this carries the caveat that it ignores thereby any potential between-diphthong correlation in F-pattern features (Rose et al. 2004). This is a problem that is still being researched.

Table 2 shows that /ɛɪ/ performs best for both different- and same-speaker comparisons, but there is not much difference between the five diphthongs. Figure 4 shows the Tippett plot for the combined LR-based discrimination on all five diphthongs. Note that the horizontal axis now runs from $\text{Log}_{10}\text{LR} = -6$ to 6, indicating that some fairly high LRs are involved for some comparisons (the LRs for non-target trials actually extend some way below $\text{Log}_{10}\text{LR} = -6$). It can be seen that once again the calibration is good, and that by combining the diphthongs the EER has decreased (from 24% for /aɪ/ on its own) to just over 10%. This is not a bad result when it is recalled that the suspect and offender samples contain values from one replicate only. Figure 4 shows that if one were to obtain a Log_{10}LR of ≥ 5 from comparison of five diphthongs,

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

for example, the chances of error (claiming strong support for the prosecution hypothesis when different speakers are actually involved) would be quite small. There were in fact two comparisons involving different speakers that had Log_{10}LRs bigger than 5 (speakers 185 & 186 with Log_{10}LR of 5.26, and speakers 182 & 199, with Log_{10}LR of 5.48. This means a probability of error of 2 in ca. 14,500, which is quite small. With smaller Log_{10}LRs the chances of error will be greater, of course.

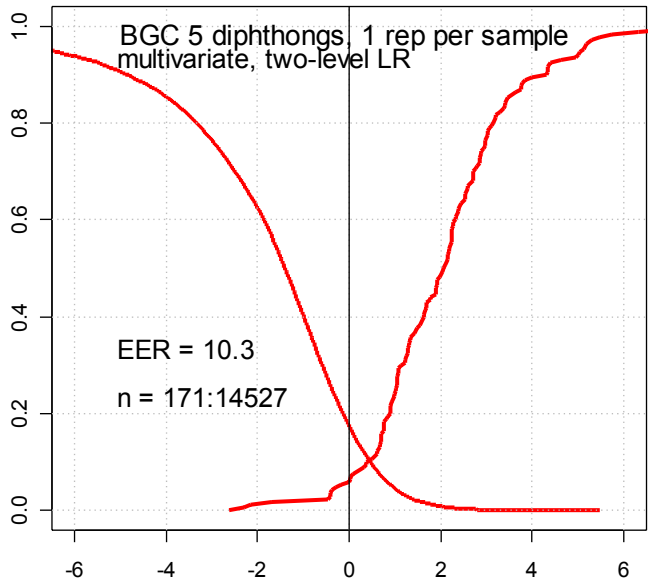

Figure 4: Tippett plot for MVLR-based discrimination using 5 Bernard diphthongs. Horizontal axis = Log_{10}LR.

Table 2: MVLR discrimination results for five diphthongs. nSS / nDS = number of target / non-target trials.

	ɛɪ	aɪ	ɐʉ	ɪə	ɛə	all 5
nSS	167	166	164	164	164	171
nDS	13,848	13,676	13,347	13,358	13,340	14,527
%corectSS	89.2	84.3	85.4	81.7	84.1	94.2
%correctDS	73.7	72.0	68.6	69.9	70.0	82.5

7. Summary & Conclusion

This paper set out to see whether diphthongal acoustics have any potential in forensic speaker recognition. Using the appropriate kernel density multivariate likelihood ratio as discriminant threshold, and setting very tough requirements of just one replicate each for suspect and offender and a pool of 171 speakers, it demonstrated with /aɪ/ that same-speaker pairs could be discriminated from different-speaker pairs reasonably well, with an EER of ca. 24%. Combining data from five different diphthongs improved the performance to an EER of ca. 10%. Thus it is clear that diphthongs have considerable potential in forensic speaker recognition, and need to be researched more.

The next step is to move to testing with more realistic data. In this experiment, only one replicate per sample was used, which is not forensically realistic: you need considerably more replicates for both suspect and offender before comparison. Furthermore, data were used from the same recording session. This would only be realistic if forensic comparison were being undertaken between suspect and offender samples separated by a short amount of time. Finally, the effect of using an intrinsic approach is unknown, and ideally an extrinsic approach should be trialed, with independent test and reference data. Non-contemporaneous diphthongal data, with several replicates per sample, must be obtained and extrinsically tested to gain a better idea of how well, and with what strength of evidence, speakers can be forensically discriminated by their diphthongal acoustics.

8. References

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

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

Bernard, J.R.L. (1967) *Some measurements of some sounds of Australian English*. Ph.D. Thesis, Sydney University.

Brümmer, N. & Du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language* Special Issue 20, 2-3, 230-275.

Ingram, J. Prandolini, R. & Ong, S. (1996) Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics* 3: 129-145.

McDougall, K. (2004) Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *Intl. J. Speech Language and the Law* 11, 1, 103-130.

McDougall, K. (2006) Dynamic features of speech and the characterisation of speakers. *Intl. J. Speech Language and the Law* 13, 1, 89-126.

Loakes, D. (2006) *A Forensic Phonetic Investigation into the Speech Patterns of Identical and Non-Identical Twins*. Ph.D. Thesis, Melbourne University.

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

Rose, P., (2006b) Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-Based Forensic Speaker Discrimination. In Berkling & Torres-Carrasquillo (Eds.), *Proc. IEEE Odyssey Speaker & Language Recognition Workshop*, IEEE, Puerto Rico.

Rose, P. Lucy, D. and Osanai, T. (2004) Linguistic-acoustic Forensic Speaker Identification with Likelihood Ratios from a Multivariate Hierarchical Random Effects Model: A 'Non-Idiot's Bayes' Approach. In Cassidy (Ed.), *Proc. 10th Australian International Conference on Speech Science and Technology* (pp. 492-497). Australian Speech Science & Technology Association, Sydney.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review