Realistic Extrinsic Forensic Speaker Discrimination with the Diphthong /aɪ/

Phil Rose[1], Yuko Kinoshita[2], Tony Alderman[3]

[1, 3] Phonetics Laboratory, Linguistics (Arts), The Australian National University.
[2] School of Languages, International Studies & Tourism, University of Canberra.
philip.rose@anu.edu.au; yuko.kinoshita@canberra.edu.au

Abstract

This paper describes a discrimination experiment in forensic speaker recognition using the Australian English diphthong /aɪ/. A two-level kernel density multivariate likelihood ratio is used as a discriminant function to investigate how well non-contemporaneous same-speaker speech samples of /aɪ/ can be forensically discriminated from different-speaker speech samples using just this diphthong's F-pattern at its two targets. Natural speech elicited from 25 Australian-English speaking males is extrinsically evaluated against a reference population of 166 male speakers from Bernard's database. Comparing samples with 12 diphthong tokens each, a respectable well-calibrated EER of between ca. 8% and 10% is obtained. Forensically important aspects of the results are discussed, including an assessment of the suitability of the reference population.

1. Introduction

This paper is another in a suite of discrimination experiments investigating how well speakers can be discriminated under forensically realistic conditions. Consistent with developments in the evaluation of forensic evidence over the last ten or so years, forensic speaker discrimination as demonstrated in this paper has a well-defined meaning. It involves, firstly, discrimination of same-speaker speech samples from different-speaker speech samples and not identification or verification of *individuals* by classic discrimination methods. Secondly, discrimination is done using a likelihood ratio as discriminant function. A likelihood ratio (LR) is the logically and legally correct way to estimate the strength of forensic identification evidence (Robertson & Vignaux 1995). In forensic speaker recognition, the LR is the ratio of the conditional probabilities of the difference between suspect and offender speech samples under competing prosecution and defence hypotheses. Consider a pair of speech samples for which it is not known whether they come from the same speaker or not. It is possible to estimate the LR for their comparison which will quantify whether the difference between them is more likely assuming same-speaker or different-speaker provenance. LRs bigger than unity indicate support for the hypothesis that the samples have come from the same speaker; LRs smaller than unity indicate support for different speaker provenance; the magnitude of the LR reflects the strength of the evidence in favour of one or other of the hypotheses. Other things being equal, the extent to which the prediction is correct is a reflection of the inherent discriminablity of the evidence.

In addition to its primary forensic use, the LR is now being recruited as a discriminant function in both automatic and traditional forensic speaker recognition, and in automatic speaker recognition in general. There have now been quite a few LR-based forensic speaker discrimination experiments, both with traditional and automatic features, which have demonstrated the viability of the LR as a discriminant function for speech (e.g. Kinoshita 2001; Rose, Osanai & Kinoshita 2003; Alderman 2005; Gonzalez-Rodriguez et al. 2006; Loakes 2006, Rose 2006). Such testing is of course crucial in the wake of the well-known *Daubert* criteria on admissibility of forensic scientific evidence, and is in line with - indeed in some cases considerably predates - the current injunction for all forensic identification science to emulate evaluation of DNA evidence by LR-based models: **'… DNA profile evidence is now seen as setting a standard for rigorous quantification of evidential weight that forensic scientists using other evidence types should seek to emulate'** (Baldwin 2005: 55).

Up to now, traditional LR-based discrimination has used mostly the formant centre-frequencies of monophthongs as features. Work has only just begun on discrimination with diphthongal F-pattern. Rose (2006) has shown for example that diphthongs have considerable discriminatory potential. However, this was with so-called *intrinsic* LR testing, where the test data also constitutes the reference population against which the two test samples are compared. Although intrinsic testing has its advantages, one of which is that the reference population is then by definition representative of the test data, it also has its drawbacks. For example, it is difficult to amass a set of test data that will be large enough to simultaneously function as a representative reference population. Furthermore, tests using independent test and reference data generally provide more realistic and defensible results. It is the aim of this paper to explore the discriminatory potential of a diphthong using extrinsic LR-based testing.

2. Data, Speakers, Elicitation, Measurement

The dataset we used is part of a larger project to find out more about the discriminability of Australian English diphthongs produced under forensically realistic conditions. Data was collected from 27 adult male AE native speakers, aged from 19 to 64 (median age 39). In this experiment, recordings from 25 speakers were used (two were discarded because their voices lacked sufficient amplitude). To be realistic, forensic discriminations need to fulfill certain criteria. Probably 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 abstract only review

most important desideratum is to allow for comparison of non-contemporaneous speech samples. This is for two reasons. Firstly, offender and suspect speech samples are usually separated by more time than is encompassed in a single recording session. Secondly, it is well-known that it is more difficult to discriminate non-contemporaneous speech samples, and therefore results from the same session would overestimate discriminant performance. Data was therefore obtained in two recording sessions, separated by between ten days and two weeks. Recordings were made using an *Edirol* R1 digital recording device at 44 kHz, coupled with a conventional external good response Sony mike, in low ambient noise surroundings, usually the recording studio at the ANU or University of Canberra.

In this experiment, we chose, for several reasons, the diphthong /aɪ/. Unlike /ɔʊ/ or /ɔɪ/ for example, both its targets tend to have three reasonably easily measureable formants, and because of this they were also well represented in the reference population, with few missing values. In /aɪ/, both F1 and F2 traverse a large part of the acoustic vowel space. /aɪ/ also tends to occur commonly in forensic speech samples:

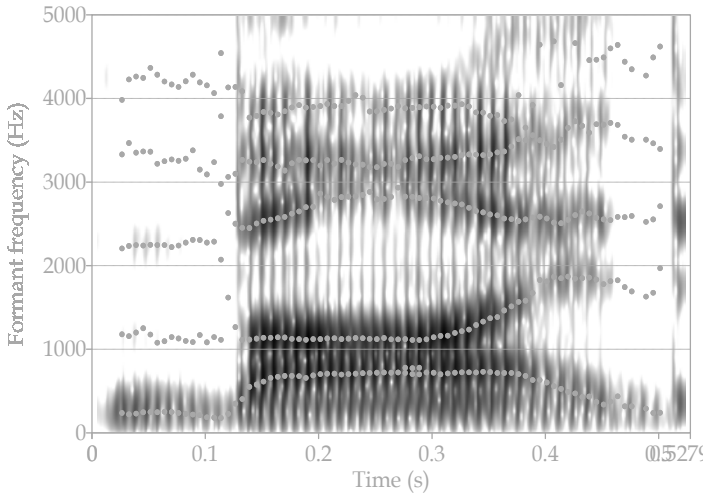

Figure 1: Spectrogram of *bide* (David 1.1) with superimposed formants showing stable F-pattern for F1 & F2 at first target.

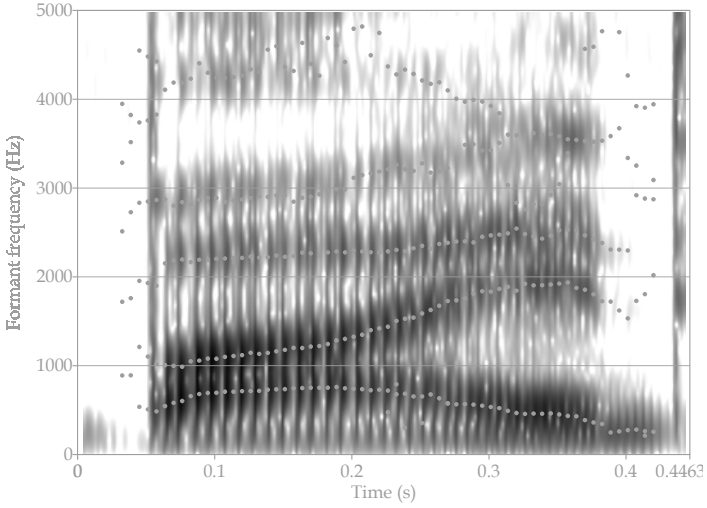

Figure 2: Spectrogram of *bide* (Jonathon 1.2) with superimposed formants showing lack of clear stable F-pattern for first target.

phone conversations often contain a 'Hi', or 'bye', for example. In contrast to other monophthongs and diphthongs, the diphthongal F-pattern of AE /aɪ/ has remained fairly stable, although there are changes in the F2 of the first target, which is now ca. 100 Hz significantly lower than 30 years ago, and in the F1 of the second target, which is now ca. 30 Hz significantly higher (Cox 1999: 24). This small amount of change was important, since we use a reference population which was recorded some 30 years ago.

In order to obtain data with some controlled consonantally induced variation in F-pattern, the corpus consisted of the six words: *buy bide high hide bite height.* It was expected that differences in F-pattern towards the end of the Nucleus would be evinced by the difference between zero, /d/ and /t/ Codas, and that differences in F-pattern at the onset of the Rhyme would be evinced by the difference between Onset consonants /h/ and /b/. The experimenter showed subjects a flash card with the target word, eg *bide*, written on it. They were asked to respond by saying the word and then spelling it out (e.g. *Bide. B I D E spells bide*). The card was removed as the speaker said the word for the first time. This method elicited of course several tokens of /aɪ/. In this paper, /aɪ/ tokens from the first word (called *readword*) and the last (*spellword*) are analysed. Two repeats of the corpus were elicited per recording session, giving 12 tokens each of readword /aɪ/ and spellword /aɪ/ in each session. Speakers typically said the two words in different intonation phrases, each word carrying the tonic accent. Common tunes were fall and rise, e.g. / [baɪd **HL**]$_{IP}$ [bi aɪ di i spɛlz baɪd **HL**]$_{IP}$ /, or / [haɪt **LH**]$_{IP}$ [hɛɪtʃ i aɪ dʒi hɛɪtʃ ti spelz haɪt **LH**]$_{IP}$ /.

Praat was used to identify and measure formants. Wideband spectrograms (0.005 sec. Gaussian window, 0 – 5k range) were generated; formants estimated with the Burg method using preemphasis from 50 Hz; and *Praat* told to find up to six formants in the range from 0 to 5k. Formant tracks were superimposed on the spectrogram, and features of their time-course used to identify points at which the diphthongal F-pattern was to be sampled.

Determination of sampling points was constrained by forensic and practical requirements, and requires some discussion and justification. In quantifying vowel acoustics, usually for linguistic-phonetic purposes, it is normal to carry over, from the auditory and articulatory realms, the notion of 'target'. The assumption is that the speaker is aiming at an auditory target – specified in terms of height, backness and rounding – and getting their supralaryngeal articulatory mechanism to do things to produce vowel acoustics that will have the appropriate perceptual consequences. Clark and Yallop (1990: 243) distinguish between stable and unstable acoustic targets. The former are characterised by invariance over time: "where the formants are parallel to the time axis" (p. 243). A nice example of this is given in the spectrogram of *bide* in figure 1, where it can be seen that the F-pattern is stable for about 10 csec. in mid Rhyme (although there is some attenuation and shifting of F3 from a nasal zero). The second target also shows a stable F2. Unstable targets are when temporal stability is not achieved because either the sound is too short, or masked by the effects of consonantally-induced perturbation (p. 244), or

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 abstract only review

Table 1: Summary statistics (mean, *between-speaker sd, mean within-speaker sd*) for /aɪ/ F-pattern in test and reference population. *n*speakers = 25 (testdata), 166 (ref. pop.); *n*replicates per speaker = 24 (test data), 2 (ref. pop.). VR = variance ratio.

	First Target			Second Target		
test data:	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**
readword	679,*44,45*	1100,*75,61*	2567,*164,113*	456,*51,80*	1940,*136,108*	2589,*130,93*
VR r-word	1.0	1.5	2.1	**0.4**	1.6	2.0
spellword	656,*53,51*	1108,*79,60*	2540,*166,125*	495,*77,113*	1839,*144,119*	2560,*150,100*
VR s-word	1.1	1.7	1.8	**0.5**	1.5	2.3
ref .data	687, *78,40*	1172, *109,44*	2473, *153,106*	444, *57,46*	1819, *142,95*	2611, *160,89*
VR ref.data	**3.8**	**6.1**	2.1	1.5	2.2	3.2

Table 2: Comparison of correlation structure between F-pattern variables in /aɪ/. Top = partial correlation in reference population; bottom = partial correlation in test data (left = readword, right = spellword).

	T1F1		**T1F2**		**T1F3**		**T2F1**		**T2F2**		**T2F3**
T1F1			**0.55**		0.16		0.21		0.25		0.08
T1F2	**0.47**	**0.62**			-0.05		-0.12		0.01		-0.06
T1F3	0.03	0.01	0.05	0.04			0.09		-0.08		**0.31**
T2F1	0.12	0.1	0.00	-0.12	-0.04	0.01			-0.18		-0.08
T2F2	0.05	0.09	0.16	0.07	0.07	0.06	-0.10	0.00			**0.35**
T2F3	0.15	-0.02	-0.14	-0.01	**0.31**	**0.42**	-0.11	-0.09	**0.56**	**0.53**	

both. The same way of thinking is used in identifying so-called tonal targets (H, L) from F0 contours in Autosegmental-metrical theory.

Now, it would be perfectly possible to identify targets of this kind in the F-pattern of the diphthongal data. However, as pointed out by Ladefoged (2003: 105), it is well known that such targets often do not line-up in time. In the F-pattern of *bide* in figure 2 for example, it can be seen that the F1 shows a short stable component a little after the rising onset perturbation associated with release of stop closure. This F1 stability is not paralleled by any stability in the F2 during this period, however, which is steadily rising, indicating that the speaker's tongue was moving forwards at the time. Thus a full specification of diphthongal targets would have to involve separate measurements for each formant, together with some indication of the time of their achievement. If we were working solely with the test data, this would be the ideal sampling strategy, and it is very likely that it would produce better discrimination results than those reported here, since speakers clearly differed in their formant dynamics. However, one of the important points of this paper is to show how, in a forensic discrimination experiment, test data must be evaluated against a suitable reference population, and in this case the reference population only has two sampling points for diphthongal targets, both at a single point in time, to represent diphthongal targets. So we were constrained to adopt a sampling strategy with only two simultaneous sampling points. In determining these sampling points, F2 was given priority. This is because in real case-work F1 (for high and possibly mid vowels) is often compromised by the lower bandpass skirt of the telephone transmission, and F3 is sometimes not well extracted (if the transmission is bad, F3 can also be compromised). The second sampling point, taken to represent the second diphthongal target (T2), was located at the point of F2 maximum, or, in the case of a stable target, over the set of stable F2 values. The first sampling point (T1) was located at the earliest point of stability in F2, where stability was interpreted as three or more extracted values with visually the same centre-frequency. In cases lacking any F2 stability, the sampling point was at the earliest point after discounting any effects from the onset consonant. Tokens with no coda (*buy, high*) often showed a stretch of either voiceless or whispery-voiced phonation at offset during which time the F-pattern continued to change (in particular F2 and F3 continued to increase). Comparison with modal tokens showed the noise-excited F-pattern values to be generally higher than in the modal tokens, and F-pattern values after end of modal phonation were ignored.

3. Processing

Likelihood ratio-based discriminations were then carried out using a two-level kernel density multivariate LR (MVLR) as discriminant function. The most important aspect of the MVLR is its ability to take correlation between variables into account, which is very important, given the potential for correlation within a diphthongal F-pattern (see Rose 2006 for further details, including the formula itself and its shortcomings. With 25 speakers, there were 25 non-contemporaneous same-speaker comparisons, where samples from each speaker's first recording session were compared with samples from their later session. 25 speakers gives 300 different-speaker comparisons. Since there were two recording sessions per speaker, a total of 1200 comparisons is possible permuting both sessions. This was reduced to 600 comparisons (i.e. two for each different-speaker pair) by comparing only same-session data (e.g. speaker 1, session 1 vs. speaker 2, session 2; speaker 1, session 2 vs. speaker 2, session 2). The MVLR formula evaluates the difference in F-pattern between the recordings for the 25 non-contemporaneous same-speaker pairs and the 600 different-speaker pairings – against a reference population. That is, a LR value is computed which estimates how much more likely the difference between the two samples under comparison is, assuming that they have come from the same speaker, given the reference population.

As mentioned above, one main point of this paper is to see what happens when the test data is evaluated extrinsically, i.e.

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 abstract only review

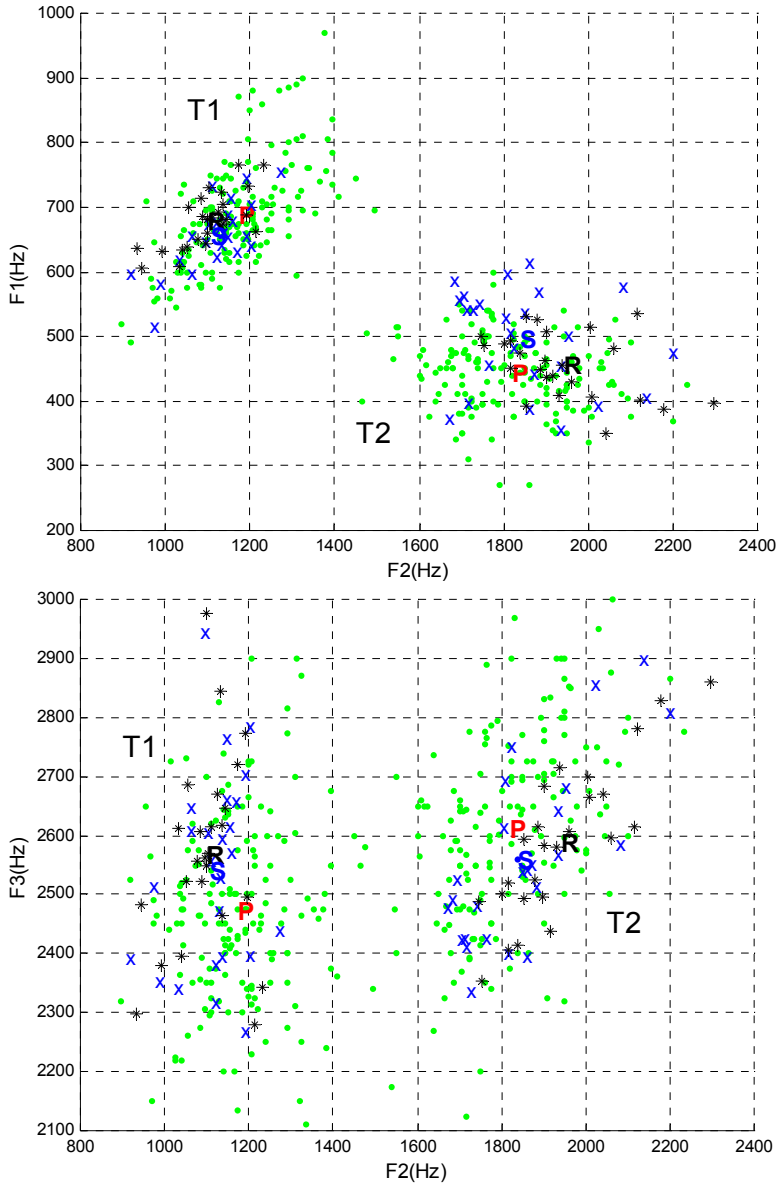

Figure 3: Comparative distribution of formant values in test data and reference population. Top = F1/F2, bottom = F2/3. T1,2 = first, second diphthongal targets. P, R, S = means of reference **P**opulation, **R**eadword, **S**pellword. Small green dots = reference population speaker means ; * = Readword speaker means; x = Spellword speaker means.

with a reference population independent of the test data. As reference population, we have used first and second target F-pattern values (F1,2,3) in /aɪ/ from 166 male speakers in Bernard's (1967) AE dataset. This contains two contemporaneous replicates per speaker of /aɪ/ in the word *hide*, one word said in isolation and one in sentence-final stressed position. Our choice is more a matter of practical necessity, as there is currently no other suitable reference population if one wants to implement the particular kernel density MVLR formula we use. A reference population – the population of all possible perpetrators excluding the suspect – is chosen primarily on the basis of the alternative hypothesis (Robertson and Vignaux 1995: 35-37). Our use of Bernard would reflect an alternative hypothesis something like: 'the voice of the offender is not my client, but is from some other adult male Australian English speaker". The Bernard data-base is actually biased towards NSW (Loakes 2006: ch. 4), but this is probably irrelevant for /aɪ/. Given the forced nature of our choice, it is important to see just how representative the Bernard /aɪ/ F-pattern data are.

Figure 3 shows the distribution of formant values in the test and reference data; table 1 summarises their important statistics. From these it can be seen that there are differences in mean values between both test and population data on the one hand, and readword and spellword conditions on the other. Two-way ANOVAs (*Speaker* vs. *Readword/Spellword*) showed that all differences between means, and interactions, are significant, most very highly so, despite the fact that some, e.g. differences at T1 for the test data, appear similar. As far as the test data are concerned, the F-pattern indicates a more peripheral articulation in the readword condition than in the spellword. The first target in /aɪ/ is very slightly backer, and more open, in the readwords than the spellwords, and the second target considerably fronter and more close. F3 in both targets is higher in readwords than spellwords. These differences are plausibly related to a pragmatic distinction between given and new. It is possible that many speakers did not feel the need to achieve quite such peripheral targets for a repeat of the same word, even when it was stressed in a separate intonational phrase. Perhaps better terms for these conditions would be *Newword* and *Givenword*.

As far as the relationship between test and reference population data is concerned, figure 3 shows that the reference population /aɪ/ has a slightly lower and fronter first target, and a closer and slightly retracted second target. The reference data have a slightly lower F3 than the test data for the first target, and a slightly higher F3 for the second. It is interesting to note that the differences between test and reference data in T1F2 and T2F1 are those that one would expect from the difference in time separating the two sets.

The similar dispersion of values in both test and reference data is noteworthy. Their principal components would probably not differ very much, although this needs to be checked with the appropriate principle component sampling distribution.

As far as correlation structure is concerned, all three sets of data are very similar. In LR-based discrimination it is the partial correlations that are of importance, since we are interested in the correlation between any two variables, given the effect of the remaining variables (Lucy 2005: 63-71). The above-diagonal part of table 2 gives the partial correlations in the reference data; below are the partial correlations in the test data, with readword values on the left and spellword values on the right. Important correlations (> 0.3) are in bold. It can be seen that all three sets of data agree in showing within-target

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 abstract only review

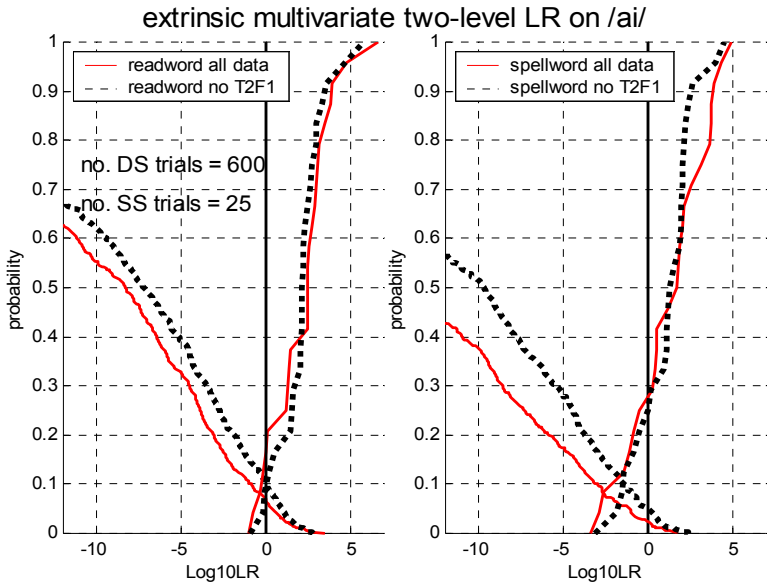

Figure 4: Tippett plots for forensic discrimination with /aɪ/ comparing use of all F-pattern information with omission of the F1 on the second target.

correlations between F1 and F2 (for T1), and between F2 and F3 (for T2); and a between-target correlation in F3. Moreover, the magnitudes of the partial correlations are also comparable.

The variance ratio data in Table 1 (between-speaker sd^2/mean within-speaker sd^2) further show that for nearly all formants the between-speaker variance is greater than the within. But not by much: in the test data the between-speaker variance only makes it once above twice the within- (and for the two T2F1 values there is greater variation within speakers than between). The variance ratios in the reference population are slightly better, but not spectacularly so, with the between-speaker variance lying between 1.5 and just over 6 times that found within speakers. These are typical variance ratio values for formants.

As far as correlation structure, variance ratio, and dispersion are concerned, then, Bernard's data set is a reasonable choice for a reference population for the test data. The differences in some mean values between test and reference data remain a worry, although it will be seen they still give good results.

4. Results

Discriminations were done separately for the readword and spellword data, and under two conditions. The first was with all available information, i.e. all three formants at both targets. In realistic circumstances, where most forensic speech samples are from phone intercepts, the F1 of high vowels is compromised by phone transmission and is of no use. It is therefore useful to know what kind of an effect, if any, this has on the LR. This, thus, was the second, 'noT2F1' condition: when F1 is omitted from the second target to simulate realistic comparison between phone recordings.

Figure 4 shows the results of these discriminations. As is now conventional in forensic speaker recognition, they are presented using a Tippett, or reliability plot, which shows cumulative distributions of LRs from same-subject trials (increasing towards the right), and different-subject trials (increasing towards the left). Results with the /aɪ/ in the readword are in the left panel; the right panel has results for the /aɪ/ in the spellword. For both readword and spellword comparisons, two curves – solid and interrupted – are shown. The solid line shows discrimination performance with all three formants at both targets. The interrupted line shows results when the F1 on T2 is not taken into account.

Forensically, the three most important aspects of the results are the discrimination performance, the strength of the evidence, and its reliability. The discrimination performance – how well the approach can discriminate between two samples from the same speaker and two samples from different speakers – is reflected in the equal error rate (EER). As can be seen, the EER for all conditions is very similar and lies between ca. 8% and 10.5%. It is lowest – ca. 8% – when readword /aɪ/s are compared using all formants, and highest – ca. 10.5% – for spellword /aɪ/s compared without their T2F1. A slight increase in the EER, from about 8% to 10%, can also be seen for both the readword and the spellword comparisons when the T2F1 is omitted. Thus some, but not much, discriminatory power is indeed lost by not being able to include T2F1.

For the spellword comparisons, there is some calibration error: $Log_{10}LR = 2.55$ for all formants, decreasing to $Log_{10}LR = 1.23$ without the T2F1. For the readword /aɪ/s, the calibration error is very small. With all formants, it is of the order of $Log_{10}LR = .39$. When T2F1 is omitted there is no calibration error: the EER is where theory predicts it to be, at the $Log_{10}LR = 0$ threshold. Thus comparisons with spellword are associated with worse calibration, and so are comparisons with all formants, as opposed to 'noT2F1'. The zero calibration error with 'noT2F1' is particularly encouraging, and is an indication that something is working perfectly. Perhaps this is related to the poor variance ratios for T2F1 (see Table 1). Another positive aspect of the results is that the calibration errors, when they occur, all favour the defence, as they should.

The role of the forensic identification expert is to estimate the strength of evidence in support of the prosecution (or defence) hypothesis. Therefore it is important to have an idea of the range of the strength of evidence to be expected with a particular set of features (here, the F-pattern at both targets in /aɪ/). In the current verbal equivalents for the strength of evidence used by the Forensic Science Service, $Log_{10}LR$s greater than 4 are characterised as "very strong". Figure 4 shows that with the readword /aɪ/ comparison, about 70% of different-speaker trials would be considered to yield *very strong* LRs. That is, for ca. 70% of different-speaker readword trials you would be *at least* 10,000 times more likely to observe the difference assuming different speakers than same speakers. This proportion increases to about 80% for the different-speaker spellword trials. Of course, for any sampling process, some different-subject pairs will differ more than others, and it is reasonable to assume that the different-speaker pairs that are resolved with sizable LRs will probably differ so much that they would not be considered suspect in the first

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 abstract only review

place. Nevertheless it is reassuring to note that such magnitudes are possible: one would not be inspired to confidence with different-speaker LRs that could not get below a $Log_{10}LR$ of, say, -1!

Figure 4 shows the same-speaker comparisons are not capable of anywhere near such strength: comparisons with readword /aɪ/ and spellword /aɪ/ have about the same upper bound of ca. $Log_{10}LR = 5$, with the readword comparisons being a little stronger. Although this counts, of course, as strong evidence, most – between ca 80% and 90% – of the same-speaker LRs can be seen to lie below $Log_{10}LR = 4$, and would not count as furnishing very strong support. This situation, at least for traditional features, appears to be typical: 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, shown in Table 1 to be poor for formants, have a limiting effect on the magnitude of the LR

It can be seen that omitting the F1 of the second target in /aɪ/ generally reduces the strength of evidence for different-speaker comparisons, where the 'noT2F1' curves appear displaced vertically relative to the curves based on all six formants. For the readword /aɪ/s, 50% of the different-speaker LRs are below about ca. -8 when based on all six formants, whereas this drops to ca. -7 if T2F1 is omitted. The loss is greater for the spellword /aɪ/s, where 50% of the different-speaker readword LRs are below -14, compared to -10 for the spellword. Omitting the T2F1 does not seem to affect the same-speaker comparisons very much. Thus some strength of evidence is lost by not being able to include T2F1, but effectively only for different-speaker comparisons.

An estimate of the strength of evidence is one thing; the reliability of the approach is another. Obviously, this is reflected in the EER, but the Tippett plots allow us to estimate reliability for any threshold. Using the results for the readword 'all formants' condition, one can see from figure 4 that with an obtained $Log_{10}LR \geq 2$, for example, there is about a 1% chance of error for the different-speaker LRs (7 of the 600 different-speaker trials were incorrectly evaluated with a $Log_{10}LR \geq 2$). LR-based comparison thus 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 $Log_{10}LR$ of 2 of more, I shall be wrong 1% of the time if I take this LR to support the defence hypothesis that the samples come from the same speaker'.

5. Summary, Conclusion, Way Ahead

The aim of this experiment was to see if useful extrinsic forensic discrimination is possible with the F-pattern in /aɪ/. Given the results, especially the EER of between 8% and 10.5%, the good calibration, and the overall strength of evidence profiles, it must be concluded that it is. As main caveats must remain the still high degree of control over the data. It is a pretty safe bet, for example, that if control is relaxed and discrimination performed between the pragmatically different non-contemporaneous readword /aɪ/s and spellword /aɪ/s, the performance will degrade considerably, at least for the same-speaker samples. The discrimination performance would also be expected to have benefited from our use of clean recordings, rather than telephone intercepts. However, given that the twin-target F-pattern thus tested does not tap all the potential information in the diphthong – it is clear from listening to the data and looking at the spectrograms that there are differences in phonation type and formant dynamics to be exploited – it must be assumed that the performance demonstrated in this paper is conservative, and a greater discriminability remains to be revealed. This will probably come from the use of the cepstrum on both targets, combined perhaps with a delta-cepstrum on the transition (although these cannot yet be extrinsically evaluated). The other diphthongs in our dataset (including /i/ and /u/ which are phonetically diphthongal for many speakers), also await evaluation.

6. References

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

Baldwin, D. J. (2005). *Weight of Evidence for Forensic DNA Profiles.* Wiley, Chichester.

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

Clark, J. & Yallop, C. (1990). *An Introduction to Phonetics and Phonology*. Blackwell, Oxford.

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

Gonzalez-Rodriguez, J. Drygajlo, A. Ramos-Castro, D. Garcia-Gomar, & M. Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language* Special Issue, 20, 2-3, 331-355.

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

Ladefoged, P. (2003). *Phonetic Data Analysis An Introduction to Fieldwork and Instrumental Techniques*. Blackwell, Oxford.

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

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists*. Wiley, Chichester.

Robertson, B. & Vignaux, T. (1995). *Interpreting Evidence*. Wiley, Chichester.

Rose, P. (2006) The Intrinsic Forensic Discriminatory Power of Diphthongs. *Proc. 11th Australasian International Conference on Speech Science & Technology*.

Rose, P. Osanai, T. & Kinoshita, Y. (2003). 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.

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 abstract only review