

Within Speaker variation in diphthongal dynamics: what can we compare?

Yuko Kinoshita¹ and Takashi Osanai²

¹Department of Communication and Education,
University of Canberra, Australia
Yuko.Kinoshita@canberra.edu.au

²National Research Institute of Police Science,
Japan,
osanai@nrrips.go.jp

Abstract

This paper studies the Australian English diphthong /aɪ/ as spoken in different speech styles. Initial visual examinations revealed that the contours of the diphthong /aɪ/ recorded for this study are affected by the speech style, and can vary considerably within speakers. This paper thus investigates the possible use of the slope of the transition of the second formant (hereafter F2), which appeared to be the most consistent feature across the context. Analysis of likelihood ratios reveal that this feature produced as good results as the F2 of the first and the second targets of / aɪ/, although it did not outperform them as we anticipated. Furthermore, this parameter appears not to have a significant correlation with the F2 of the first target. This allows us to combine those two parameters and produce a stronger result.

1. Introduction

Since its appearance in 2001, forensic speaker identification research based on the likelihood ratio (henceforth LR) has been slowly but surely growing. Kinoshita (2001) explored this LR-based methodology from a linguistic point of view, using the formants of five vowels and several consonants in Japanese. Around the same time Meuwly & Drygajlo (2001) incorporated LRs in a speech engineering approach. Although these are studies produced from two different sides of speech research, the ultimate aims of those studies were the same – evaluating speech evidence in an objective and systematic manner and presenting it to the court in a legally appropriate fashion. The LR-based approach is a scientifically correct way to present evidence in court. However, in order to comply with the Daubert ruling in the US Supreme Court (available from <http://supct.law.cornell.edu/supct/html/92-102.ZS.html>), which set standards for the admissibility of evidence, the LR-based approach must be further researched to bridge the gap in the knowledge of error rates produced by the commonly used parameters.

The LR-based linguistic phonetic approach has so far been applied to some of the commonly used parameters, such as the formants and cepstrum of monophthongs and some consonants, in Australian English and Japanese. As well as the above mentioned Kinoshita (2001), Alderman (2004) explored the application of the LR-based method to Australian monophthongs, using Bernard (1970)'s data as the population data. Rose, Osanai, & Kinoshita (2003) applied the same approach to both cepstrum and formants extracted from Japanese telephone speech. Diphthongs however remain to be researched to date: this study thus aims to report a preliminary study of the application of the LR-based approach to Australian English diphthongs.

Although diphthongs have not yet been researched using LR-based methods, there is forensic speaker identification research on these vowels. McDougall (2005) investigated the possibility of the use of the formant dynamics of Australian English /aɪ/ diphthongs followed by a consonant /k/ as a

parameter to classify speakers, based on Discriminant Analysis. The utterances used in this study were well controlled (readout speech recorded in a single recording session) and the population was small (five speakers), and it is thus difficult to refer to its results in court. McDougall's study, however, showed considerable promise in the use of diphthongs and provides a good ground for the further investigation of Australian English diphthongs under forensically more realistic conditions.

One of the problems that experts face in forensic speaker identification is that typically we do not have any control over the speech data that we obtain. The recording equipment used to record the incriminating speech and the suspect's speech will be most likely to be different in real forensic cases. Automatic speaker identification parameters — such as cepstrum coefficients — perform better than the formants commonly used in linguistic phonetics (for instance, see Rose, Osanai, & Kinoshita 2003), but are still quite susceptible to these discrepancies in the recording equipments and/or quality. There is research attempting to overcome this problem and we see some promise (for instance, see Alexander, Dessimoz, Botti, & Drygajlo, 2005). However, for now at least, this provides a strong reason for the researching the traditional acoustic parameters, such as formants, which seem to be more robust against non-matching recording conditions. This paper thus chooses to research the use of formants in diphthongs.

Diphthongs appear to be a useful target for forensic speaker identification. However, we found that the diphthong /aɪ/ reveals rather large style-dependent within-speaker variation in its contour of F2 during our initial examinations of the formant contours. These variations are so large that the conventional comparison methods for diphthongs, such as comparing the formants of the first and the second targets or glide, will simply not work when the utterances to be compared were produced in different speech styles (see Clermont 1991:51-3) for a comprehensive summary of previous studies.)

For instance, Figure 1 below shows the formant contours of two speakers, AW and DHI, with 12 utterances of /aɪ/.

These were spoken under three different conditions, and hence in different speech styles. In this example, the contours of F2 can be clearly divided into two groups, using three criteria: duration, the frequency at the first target of the diphthong, and the shape of the contours. Two of the three speech styles fall into one of these groups, and the other style corresponds to the other group.

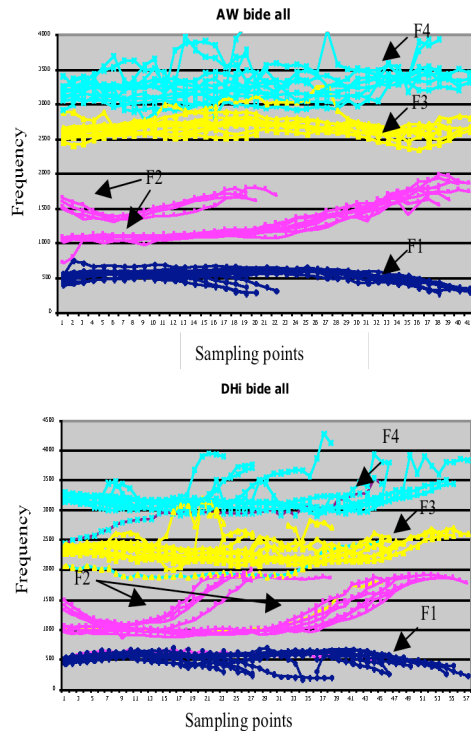


Figure 1: Speaker AW and DHI's formant contours for 12 utterances of /aɪ/.

Figure 1 suggests not only that the styles affect the shape of the formant contours but also that the effects of the styles seem to differ from speaker to speaker. This presents us with a very fundamental question: “what is comparable?”

This paper thus investigates the usability of the traditional measuring points, and also searches for more robust parameters for characterising individuals.

2. Data

2.1. Dataset collection

The dataset was collected for a larger project that aims to build more knowledge on the diphthongs produced under forensically realistic conditions. We used 27 male adult native speakers of Australian English, aged from 19 to 64 (average age was approximately 39). The recordings were made mostly at the recording studios at University of Canberra or The Australian National University, but some of them were made at other quiet locations, such as the author's home or the offices of the participants.

Recordings were made with an Edirol R1, recorded as .wav file at a sampling rate at 44,000 Hz with an external microphone. All speakers had two recording sessions separated by 10 days or more — most of them were separated by about two weeks.

Recording natural utterances was important, but it was equally important to elicit a sufficient number of target

diphthongs. We therefore employed the following recording technique.

Speakers were presented with a card, on which one word was written. First, they were asked to read out the word. Then the card was hidden from them, and they were asked to spell the word that they just read out using a set sentence structure: “X-X-X-X spells XXXX”. For instance, when the card ‘bide’ was presented, the speakers should say “Bide. B-i-d-e spells bide.”

The recording included 35 words: 34 target words and one dummy at the beginning of the recording. Seven diphthongs were recorded: /aɪ/, /aʊ/, /eɪ/, /ɔɪ/, /oʊ/, /ɪə/, and /ɛə/; in the h_d, h_t, b_d, b_t, h_#, and b_# contexts. Not all diphthongs can be embedded in this frame and make words, so some combinations were not recorded. Some unusual words had to be included in order to elicit the target vowels, and they were presented as proper nouns, so that the speakers were not distracted by meaningless words. Table 1 below lists the words used in the recording.

Table 1: List of words included in the recording

	h_t / h_d	b_t / b_d	h_# / b_#
/aɪ/	height / hide	bite / bide	high / buy
/aʊ/	how'd	bout / bowed	how / bough
/eɪ/	hate / Haydes	bait / spade	hay / bay
/ɔɪ/	Hoytes	buoyed	boy / coy
/oʊ/	Hote / hoed	boat bode	hoe / bow
/ɪə/	adhered	beard	hear / beer
/ɛə/	haired	bared	hair / bear

2.2. Data selection for analysis

In this study we randomly selected 10 speakers from the 27 in the full dataset for analysis. Since our aim is finding an appropriate parameter for the larger study, the analysis was limited to one word: *bide*. The formant contours of /aɪ/ in three different contexts per utterance were sampled:

- 1) read out word ‘*bide*’,
- 2) ‘i’ in spelling of the word ‘*b-i-d-e*’,
- 3) sentence final position of ‘*b-i-d-e spells bide*.’

The diphthong in the first context is the most controlled. The second context is considered to be the closest to natural speech, since the speakers utter this referring to their memory of the word. The third context is uttered based on a given frame, so it is not a completely spontaneous utterance. However it would be more natural compared to the read out word, considering that the speakers utter the diphthong relying on their memory. Also this one does not carry the focus of the sentence, so it is assumed to be less carefully articulated.

Hereafter the three styles — produced from these three contexts — will be referred to as “Word”, “Spelling”, and “Sentences”.

We anticipated discrimination to be best when two “Word” style samples were compared, then followed by the “Sentences” style, and finally the “Spelling” style being poorest. We also anticipated discrimination to be poor using tokens uttered in different styles, especially when the combination involves the “Spelling” style.

3. Procedure

3.1. What is comparable?

Characterising the formant structures of monophthongs is a relatively straightforward task. Most commonly, the midpoint of the vowel duration or the mean of the stable section near the middle of the vowel duration is used, and these selected points are usually available across the speakers and different utterances. However, things become much more complicated with diphthongs. The traditional approach is use set sampling points, such as onset, the first target, the second target, and offset. However, as shown in the previous section, visual inspections of the formant contours of ten speakers' utterance of *bide* suggested that the traditional sampling points may not be useful for forensic speaker identification.

McDougall (2005) reports that speakers are better characterised using the dynamic features of the formant transition rather than a small time slice. She proposes an approach that uses polynomial linear regression to model the formant contour of the entire diphthong and reports promising results. Observation of the data for the current study, however, suggests the use of the shape of the whole contour will not be effective when the diphthong was produced in different contexts and/or more spontaneous speech styles.

There was however one feature that appeared to be relatively stable across the contexts: the slope of F2 during the glide. As F2 is well within the range of the telephone bandwidth and the formant tracking of F2 is relatively reliable, this is an attractive option as an additional parameter. Thus the current study applies the LR-based approach to this potential parameter, as well as the traditional two targets, and examines its performance in speaker discrimination.

3.2. Measurements of the targets

3.2.1. Target 1 and 2

We evaluated the usefulness of the slope of F2 by comparing its speaker identification ability to that of the traditional parameters, the targets of the diphthong. Diphthongs by definition have two targets within them, but those two targets are not always clearly realised. In this study, the first target was sampled by measuring the formant in the region where F1 and F2 are most stable. It was however not always possible to define the first target by this method — in quite a few cases, F1 never achieved a stable state. Also in some cases F1 stabilised at a different place from F2. In these cases, the first target was sampled by measuring the formant in the region where F2 stabilised.

The second target was even more difficult to define, as most (but not all) speakers never achieved a steady state for the second target. In those cases, the highest point of F2 was considered as the second target.

These measuring points are henceforth referred to as T1 and T2.

3.2.2. Slope of the glide

The slope of F2 in the glide was measured using linear trend line fitting. The beginning and the end points of the glides were picked by visual inspection of the contours of F2: from the end of the stable part of the first target, to the beginning of the stable part of the second target. Then the linear trend line was fitted to the slope. Since the coefficients obtained from this process represent the angle of the slope, those coefficients

were used as the parameter that evaluates the slope of F2 — hereafter called “G”.

Figure 2 below is an example of the calculation of the coefficients. This is the slope which was extracted from the entire duration of /ai/. “ $y=6289.5x-11082$ ” shown in the figure is the expression of the trend line which was fitted to the slope of the glide. The coefficient 6289.5 (Hz) was then saved as G for this sample.

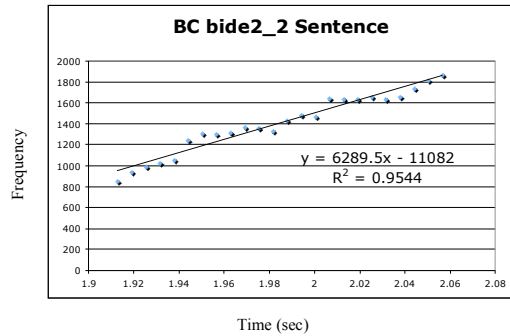


Figure 2: An example of the sampling of the coefficients.

In Table 2, the results of the measurements and calculations of the three parameters are summarised: T1, T2 and G, together with their means, standard deviations and coefficients of variation.

Table 2: The result of the measurements

		“Word”	“Spelling”	“Sentence”
T1	Mean	1087	1257	1098
	SD	28.29	73.52	36.08
	C of variation	2.6	5.85	3.29
G	Mean	5467	7040	5376
	SD	842.8	1165.43	904.5
	C of variation	15.42	16.55	16.83
T2	Mean	1872	1926	1785
	SD	61.07	95.37	100.13
	C of variation	3.26	4.95	5.61

The three different speech styles specified previously were presented separately. Target 1 in the “Spelling” style produced notably higher T1 than the other two styles. This is probably more to do with the influence of the preceding vowel /i/ rather than the speech style itself. The diphthongs in question were produced after the consonant /b/ for “Word” and “Sentence” styles, but it followed a high front /i/ when it was spelled out (ie. [bi: ar di: i:]).

Also the comparisons of the coefficients of variation shows that “Word” style produces the smallest variation across all the measuring points, whereas “Spelling” and “Sentence” styles produce similarly sized variation, except for where T1 was sampled.

In order to use the coefficients in the LR calculation, their distribution needs to be examined, because the formula used for the LR calculation assumes the normality of the distribution of the data.

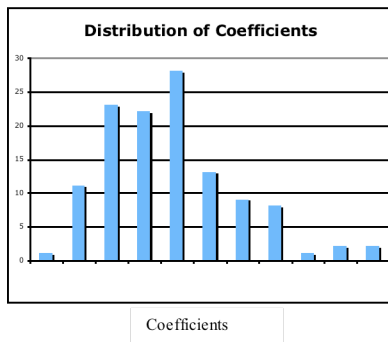


Figure 3: The distribution of the coefficients.

Considering the small size of the population, the distribution of the coefficients is close enough to normal. This study thus proceeds to employ Lindley's LR calculation formula, which is described in detail in the following section.

3.2.3. Application of Lindley's formula

The LR is the probability that the evidence would occur if the assertion is true, divided by the probability that the evidence would occur if the assertion is not true (Robertson & Vignaux, 1995:17). In the forensic speaker identification context, it will be the probability of observing the evidence if it came from the same speaker (usually the prosecution hypothesis) divided by the probability of observing the same evidence if it was produced by different individuals (usually the defense hypothesis). When P represents probability, E represents evidence, and H stands for hypothesis, this can be expressed as follows:

$$LR = \frac{P(E|H)}{P(E|\bar{H})} \quad (1)$$

The LR is always expressed as a positive ratio. The LR will be larger than 1 when the given evidence supports the hypothesis, and smaller than 1 when the evidence does not support the hypothesis. The relative distance of the LR from 1 measures the strength of the evidence.

It is important to note here that the relationship of the LR to 1 determines which hypothesis the piece of evidence supports, the LR is not a binary expression of truth. In other words, it does not answer the question "Are these two samples from the same speaker?" but rather expresses the strength of evidence on a continuous scale.

The production of the LR calls for the probabilities of observing the given evidence when each of two opposing hypotheses is true. When the data is categorical, such as gender or age, this calculation is straightforward. For instance, if we have 100 guests at a party, 65 male and 35 female, the probability of a male guest winning a door prize is 65%, and the probability of a male guest not winning a door prize is 35%, leading us to a LR 1.86.

Most speech evidence, however, cannot be processed in this manner. Speech data based on measurements of acoustic parameters are continuous, and the calculation of those probabilities with continuous data requires mathematically more complex modelling. With continuous data, where f represents the probability density of the evidence E, the simple formula for LR (1) becomes:

$$LR = \frac{f(E|H)}{f(E|\bar{H})} \quad (2)$$

In order to produce the necessary probability densities, Lindley's formula (quoted from Aitken 1995:181) takes seven values into account. Those values are:

Mean of criminal samples (\bar{x}),

Mean of suspect samples (\bar{y}),

Number of measurements of criminal samples (m),

Number of measurements of suspect samples (n),

Variance of criminal and suspects samples (σ^2),

Overall mean of population (μ),

Overall variance of population (τ^2).

Furthermore, z , w , and a^2 are derived using the values above as:

$$z = (\bar{x} + \bar{y}) / 2, w = (m\bar{x} + n\bar{y}) / (m + n), \quad (2)$$

$$a^2 = 1 / m + 1 / n.$$

Using all the values above, and also assuming that all the data are distributed normally, Lindley developed the following formula, in which V is the estimate of the LR for the evidence:

$$V \equiv \frac{\tau}{a\sigma} \times \exp\left[-\frac{(\bar{x} - \bar{y})^2}{2a^2\sigma^2}\right] \times \exp\left[-\frac{(w - \mu)^2}{2\tau^2} + \frac{(z - \mu)^2}{\tau^2}\right] \quad (3)$$

Two important factors in this formula are similarity and typicality. Given that we are comparing two sets of data, the similarity between them is clearly of interest. However, the similarity by itself is not sufficient to evaluate the evidence. Typicality also needs to be taken into account. If the two sets of data were located centrally in the population distribution, the significance of their similarity decreases, as high typicality means that there are many others who could produce the data in question. In this formula, similarity and typicality are expressed in the second and third items respectively, and the first item shows how much larger the SD of the population is compared to the SD of the testing data (Rose 2002:311).

It should be noted that V is not the same as the LR: what this formula produces is not an LR, but an estimate of it. Earlier in this section, an example of categorical data was presented, together with the LR calculation for it. In that particular scenario, it was indeed possible to produce an actual LR. It was known that there were 65 male and 35 female guests. We therefore can produce true probabilities of a male guest winning a prize and not winning a prize based on the known facts. In the case of continuous data, however, the calculation relies on a mathematical model of reality, rather than reality per se. It is thus not possible to obtain a true LR from the formula.

3.3. Comparisons

This study involved 10 speakers and each of them was recorded on two separate occasions. They also repeated the task at each recording session. Since they produced /aɪ/ in three different contexts, all speakers produced 12 /aɪ/ each (two utterances * two recording sessions * three styles).

Since this is a very small dataset, the population data was estimated using a cross validation approach. That is, a pair of speakers are removed from the population and used as the test sample. The rest becomes a reference sample. This is repeated, replacing one speaker after another out of the population until all possible speaker combinations have been taken out from the reference population and tested. In this way, a test sample and a reference population can be independent. Also repeating the tests with many different test sets reduces the chance of being strongly biased by a peculiar piece of data which happened to be the test sample.

Since there were ten speakers recorded on two occasions, the test produced 180 different-speaker comparisons. As for the same-speaker comparisons, there are only ten possible combinations, but since they were evaluated with nine

different populations because of the cross validation approach, 90 calculations were made altogether.

4. Result

4.1. Result of estimated LR calculation

4.1.1. LogLR

In the expression of the LR, it is common practice to use logarithmic scaling. In the logarithmic scale, the likelihood ratio is expressed from negative values to positive, having 0 as the neutral, not 1. In other words, the evidence with LR 10 holds equal strength as that with -10, though one supports the prosecution hypothesis and the other supports the defence hypothesis. This is easier for a layperson to grasp intuitively than the system in which 10 and 0.1 has equal strength.

Furthermore, some combinations of speech data used in this study produced such small LR estimates that they could not be further processed by the computers used. The logarithmic scale was useful in this situation. Thus this paper uses logarithmic scale of the LR estimates (logLRs) hereafter.

4.1.2. Correlations

In order to combine multiple parameters using Bayes theorem, the combined parameters must not have correlations. It is expected that formants will have some correlations among them given they are produced using the same articulators. However, testing the first four formants of five vowels and the consonants /m/ and /s/ in Japanese in (Kinoshita, 2001) revealed that there is very little correlation among those parameters. In this study as well, very little correlation among the three parameters was found. Table 3 summarises the results of the correlation tests. The columns 'T1' and 'T2' show results where the first and the second target respectively were used as the parameter, and 'G' shows results where the coefficients of the slope were used.

Table 3: Correlation between parameters

T1/G	T1/T2	G/T2
-0.020	0.253	0.175

It is thus assumed that the three parameters studied in this paper do not have significant correlation and thus it is valid to combine the LRs of all three parameters in the classification of the speakers.

4.1.3. Equal error rate (EER)

As has been described above, LR is generally interpreted in its relation to 1 (or 0 in case of logLRs). However, in this study what was obtained from Lindley's formula is used as a discriminant function, rather than an LR in a real sense. In this case, the use of 1 or 0 as a threshold becomes a rather meaningless exercise. Thus, we discuss the results in terms of its equal error rate (henceforth EER). EER shows the error rate at the place where the two distributions in question are separated from each other most effectively. In this paper the two distributions are the distribution of logLR of within-speaker comparisons and that of between-speaker comparisons. Although EER is not a useful measure to present the evaluation of the evidence in court, it is a useful tool to compare the performance of the methodologies or the selected parameters.

In addition to calculating EERs for the LRs for each of the three parameters T1, T2, and G, we also produced the

EER of the combined LRs. Having found very small correlations among the three parameters, we proceed to combine LRs from all them using Bayes' theorem – ie. adding up the logLRs to produce the combined logLR.

Tables 4 and 5 below present these EER results. Each row gives results comparing two of the three styles. 'W', 'Sp', and 'S' mean "Word", "Spelling", and "Sentence" respectively. Therefore W/W means that data from the "Word" style was compared to another dataset from the "Word" style, whereas W/Sp means that data from the "Word" style was compared to data from the "Spelling" style — ie, when the speaker spelled out the word.

Table 4 presents the results where each parameter was compared separately: the columns 'T1' and 'T2' show results where the first and the second targets respectively were used as the parameter, and 'G' shows results where the coefficients of the slope were used.

Table 5 shows the performance where the three parameters were combined. For instance, 'T1+G' shows the EERs where the results from T1 and G were combined. 'All' shows the results where all three parameters were combined.

Table 4: EERs for the three parameters with different speech styles.

EER produced with each parameter

	T1	Glide	T2
W/W	29.44%	31.67%	33.89%
Sp/Sp	31.11%	32.22%	35.83%
S/S	34.72%	26.94%	37.78%
W/Sp	52.64%	50.69%	38.47%
W/S	26.39%	36.11%	44.44%
Sp/S	60.56%	46.39%	47.22%
Combined	27.78%	22.78%	26.94%
Mean	37.52%	35.26%	37.80%

Table 5: EERs for the three parameters with different combination of speech styles.

EER produced by combinations of the parameters

	T1+G	T1+T2	G+T2	All
W/W	13.06%	32.22%	20.56%	9.17%
Sp/Sp	15.56%	35.28%	24.44%	11.67%
S/S	16.11%	27.78%	30.00%	10.28%
W/Sp	22.22%	31.39%	34.58%	20.83%
W/S	13.89%	31.94%	32.78%	14.03%
Sp/S	35.00%	39.72%	48.89%	30.00%
Combined	1.94%	25.83%	12.78%	0.00%
Mean	16.83%	32.02%	29.15%	13.71%

5. Discussion

The initial visual inspections of the formant contours suggested that the shape of the slope during the transition of targets would be less susceptible against differences in speech style than the target itself. It was thus anticipated that the parameter 'Glide' would outperform the other two parameters, especially with the comparison where different speech styles were compared. However, the result did not show such tendency. Although it is difficult to tell from this small set of data, the overall angle of the slope of the glide does not perform any better than the two targets of the diphthong.

However, this coefficient seems to be at least as useful as Target 1 and 2 as a speaker identification parameter. In the

LR-based approach, the LRs obtained from multiple parameters can be combined using Baye's theorem. With one parameter, the EERs on average ranged 35.26% to 37.80%, with two parameters 16.83% to 32.02% and with all three together the mean EER was 13.71%. Considering that the EER improved clearly as we add extra parameters, the discovery of another parameter is a very useful contribution by itself. Furthermore, the combination of Target 1 and Glide performs much better (EER 16.83%) than the other combinations (32.02% and 29.15%) where two parameters were used, which suggests the possibility of the slope of the glide carrying information which complements the information that the target carries.

This study employed a simple measure to incorporate the glide into forensic speaker identification. With refinement of the method, such as breaking the slope into a few sections so that it reflects the shape as well as slope of the glide, perhaps some improvement in its discriminatory power can be observed.

6. Conclusion

This paper investigated the use of the slope of F2 in the glide of the diphthong /ai/. Contrary to expectations, the angle of the slope of F2 was not found to be particularly robust against differences in speech styles. However, the angle of the glide was at least as useful as the two targets of the diphthongs, which are traditionally measured in the discussion of the nature of the diphthongs. Considering this paper took a rather simple approach to characterising the glide, this is a promising result. Further investigation using a method which allows us to characterise the shape of the slope as well as the angle may improve its performance. Also, the results seem to suggest that this pilot study is worth extending. Thus further experiments with more speakers will be the next step.

7. References

- Aitken, C. G. G. (1995) *Statistics and the Evaluation of Evidence for Forensic Scientists*. New York: Wiley.
- Alderman, T. B. (2004). Refining the likelihood ratio approach to forensic speaker identification: The effects of non-normality in the background distribution as modelled with the Bernard data for Australian English. Unpublished Honours thesis, The Australian National University, Canberra.
- Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *The International Journal of Speech, Language and the Law*, 12(2), 214-234.
- Bernard, J. (1970). Toward the acoustic specification of Australian English. *Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 23(2-3) 113-128.
- Clermont, F. (1991). Formant-contour models of diphthongs: a study in acoustic phonetics and computer modelling of speech. Unpublished PhD thesis, The Australian National University, Canberra.
- Kinoshita, Y. (2001). Testing realistic forensic speaker identification in Japanese: A likelihood ratio based approach using formants. Unpublished PhD thesis, The Australian National University, Canberra.
- McDougall, K. (2005). The role of formant dynamics in determining speaker identity. Unpublished PhD thesis, University of Cambridge, Cambridge.
- Meuwly, D., & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian framework. *2001: A Speaker Odyssey. The Speaker Recognition Workshop*, Crete, Greece, 145-150.
- Robertson, B., & Vignaux, G. A. (1995). *Interepreting Evidence*. Chichester: Wiley.
- Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.
- 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. *The International Journal of Speech, Language and the Law* 10 (2), 179-202.