Likelihood Ratio-based Forensic Voice Comparison with the Cantonese Diphthong /ei/ F-pattern

Pang Jialin[1] *& Phil Rose*[1,2]

[1] Division of Humanities, School of Humanities and Social Science, Hong Kong University of Science and Technology, Hong Kong, China
[2] School of Language Studies, Australian National University
jpangaa@ust.hk, philip.rose@anu.edu.au

Abstract

This paper describes a likelihood ratio-based forensic voice comparison experiment using acoustics from the formant pattern of the Cantonese diphthong /ei/ in the word /sei/ "4". Non-contemporaneous recordings from 18 young male native speakers are used, each containing nine replicates. Multivariate likelihood ratios derived from the cubic polynomial coefficients of the /ei/ F-pattern are obtained with a two-level kernel density multivariate likelihood ratio. A calibrated cross-validated log-likelihood ratio cost of 0.46 is obtained, with an equal error rate of about 14%. It is concluded that the /ei/ diphthongal F-pattern can be useful in forensic speaker identification in Cantonese.

Index Terms: Forensic voice comparison, likelihood ratio, Cantonese, diphthong.

1. Introduction

Over the last twenty years, an increasing number of experts have paid attention to the proper evaluation of forensic evidence. This has been mainly the result of the successful use of forensic DNA profiling and its way of evaluating evidence, but more recently forensic voice comparison has also made a contribution [1]. As claimed in [2] "...*the* fundamental formula of forensic science interpretation" is Bayes' Theorem, which states that the odds in favour of a hypothesis, given the evidence brought in its support, is the product of the strength of the evidence and the odds in favour of the hypothesis before the evidence is adduced.

The Likelihood Ratio of Bayes' Theorem has been used in forensic voice comparison to test the efficacy of various sounds. Diphthongs have been studied quite a lot within this approach, e.g [3, 4, 5], as their acoustics promise to contain a lot of speaker-dependent information. In Chinese, however, only three sounds have been studied so far: two monophthongs [i, y] [6] and a triphthong [iau] [7], and only in one variety – Shenyang Mandarin. Up till now, no studies have been done on Cantonese, so this paper looks at forensic voice comparison with a Cantonese diphthong /ei/.

2. Procedure

2.1. Corpus & elicitation

The suitability for FVC of the Cantonese diphthong /ei/ was tested with the speech of Cantonese males from the Hong Kong MTR forensic database. This is a small, quasi map-task database primarily designed to elicit natural, unreflected, but controlled speech for testing likelihood ratio-based approaches to forensic voice comparison. Subjects were given a map of the Hong Kong Mass Transit Railway and were asked various questions about it, for example how to get from station x to station y, or how many stations there are between station a and station b. For this paper we used the response to the second type of question, where the expected number of stations was four (in Cantonese: /sei 33/ 四). Speakers were encouraged to give answers repeating the question, and to count out loud if they wanted, thus a typical question-response was:

> Q: *Taaigú tòhngmàih Tòhnglòhwaān jīgaān yáuh géidōgo jaahm a?*
> A: *Taaigú tòhngmàih Tòhnglòhwaān jīgaān ... yáuh seigo jaahm.*
> Q: *How many stations are there between Taiku and Causeway Bay?*
> A: *Between Taiku and Causeway Bay there are ... four stations.*

Before recording the speaker was allowed to practice to make sure that they counted in the way we wanted (different people understand *how many stations between x and y* differently – some include x and y, for example). During the recording, if a speaker forgot to give a full answer, they were prompted. The database was designed to elicit up to 10 replicates of /sei/ per recording session, although if speakers counted-out loud (some counted *sotto voce*), there were usually several more. It is an essential component of a forensic voice comparison database to include non-contemporaneous recordings [8]. For this experiment speakers were recorded on two occasions separated by at least a month. The questions asked in the two recording sessions differed, to try and avoid familiarity effects, but it is likely that the participants did not remember very much of the first set of questions over the space of a month. Generally there were no problems in getting natural-sounding speech, and many tokens of /sei/, this way.

2.2. Speakers

Our recordings were collected from 18 male native Cantonese speakers. Most were native Hong Kong citizens, with the remainder from Guangzhou. They were aged from 22 to 52 and were all students of the Science and Technology University of Hong Kong (HKUST). None of them had linguistic educational background.

2.3. Recording

Recording was carried out in a quiet room on the HKUST campus. In order to simulate a phone conversation, the speaker was located in a separate room from the researcher, who contacted him by mobile phone. Recording was done by the researcher's assistant, who was in the room together with the

speaker, using a Sony ECM-MS907 microphone and a notebook computer with *Cool Edit* software. Recordings were made at 44100 Hz with 16 bit resolution and saved as .wav files.

2.4. Measurement & examples of between-speaker variation

Using *Praat*, utterances containing /sei/ were located in the recordings, edited-out and saved. We used nine replicates per speaker per recording. The acoustic measurements were then made from these edited files. Early work on diphthongal acoustics within the likelihood ratio paradigm, e.g. [4, 5] had sampled formant values at given putative target points. In [9] and [10]] work was reported trialling the idea of using formant trajectories for speaker characterisation, still, however sampling formants at discrete points. However in [3] it was shown that stronger evidence could be achieved by actually quantifying formant trajectories with coefficients of their polynomial or DCT time functions, rather than as in the previous work, and this is the approach we adopt here.

This approach has its drawbacks, however, as formant trajectories are often quite strongly differentially influenced by neighbouring sounds. In our case, all /ei/ tokens were preceded by an effectively invariant /s/, so its effect on the initial F-pattern is controlled for. All tokens were also followed by the /k/ in the general classifier *go* 個, but in this case there was considerable between- and within-speaker variation in its realisation. Usually some lenition was observed for this consonant, and realisations from [g] to [ɣ] through zero were common. In the latter case, of course, there is no discrete point at which the /ei/ F-pattern trajectory can be said to end and the /ɔ/ F-pattern begin, so the best approach is to chose a sampling strategy and apply it consistently. Our strategy was to adjudge as offset the F2 maximum. Onset was not problematic and was taken to be at the first strong glottal pulse.

Figures 1 to 3 show some examples of between-speaker variation in /ei/ F-pattern. Figure 1 shows a typical spectrogram of a token of /sei/ in seigo *four*, with the /k/ lenited to [ɣ]. An expected F-pattern is seen for a diphthong with a mid front unrounded first target and a high front unrounded second target. Both F2 and F3 show the consonantally induced perturbations onsetting at about the same time.

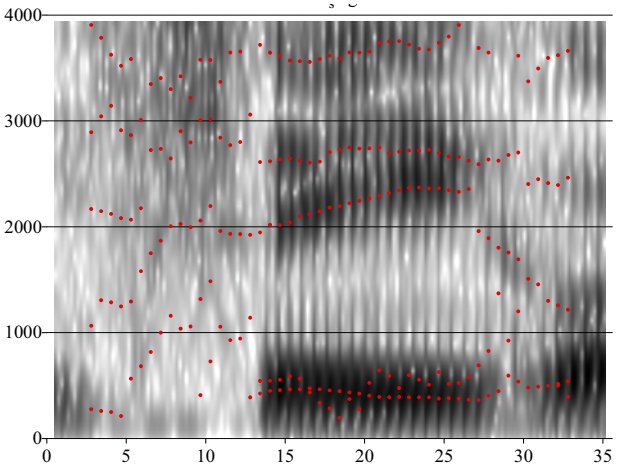

Figure 1: *Spectrogram of /sei/ in seigo showing* /k/ lenited to [ɣ].

Figure 2 shows a spectrogram of /sei/ from another speaker, also with lenited /k/, where the F3 perturbation appears to start well before that of the F2, which reaches its maximum very late. Figure 3 shows a spectrogram of a speaker who has completely lenited his /k/ and glides from [i] to [ɔ]. In this token, the F3 perturbation also starts before the F2 maximum is reached.

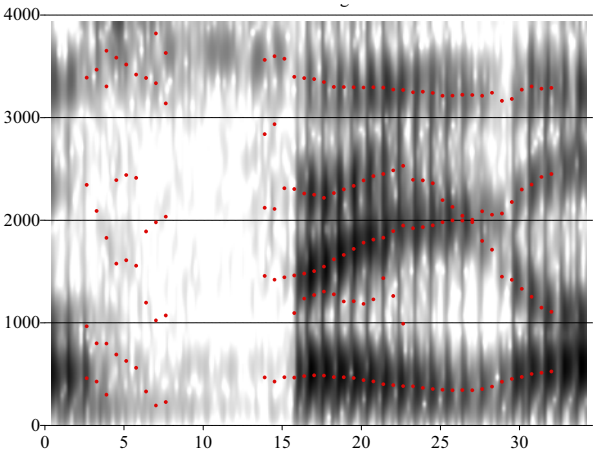

Figure 2: *Spectrogram of /sei/ in seigo showing* /k/ lenited to [ɣ] and early F3 perturbation.

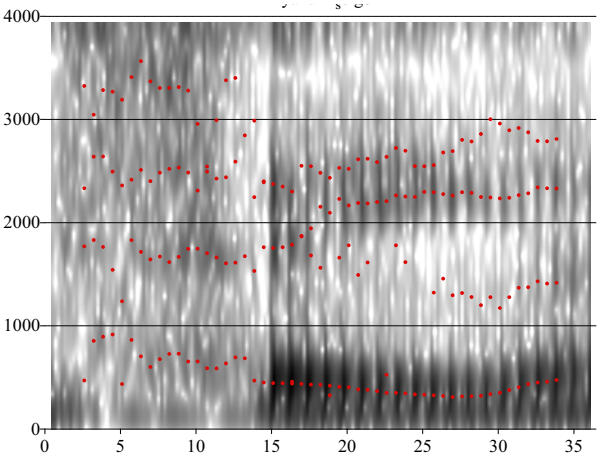

Figure 3: *Spectrogram of /sei/ in seigo showing* /k/ lenited to zero and early F3 perturbation.

We used a default setting of Burg, with five formants below 4 kHz and the *formant listing* function of *Praat*, to save the resulting F-pattern centre-frequencies from adjudged onset to offset, and their corresponding duration measurements, in Excel. (One expects four formants from males for this sound below 4 k, and it sensible to allow for an additional extraneous pole from other, e.g. subglottal, sources.) There is no setting for formant extraction that will work with every speaker, nor indeed one that will work always with a single speaker. If the setting did not produce good-looking tracking against the wide-band spectrogram, different settings were tried and the best looking one chosen. Obviously incorrect values were manually corrected.

2.5. Parametrisation, examples of within-speaker variation, and further processing.

Following [3], code was written in *R* to model the F-pattern trajectories with cubic polynomials. The non-

contemporaneous within-speaker variation was typical: some speakers did not differ much; some differed in F2; some in F3 and some in both. Figure 4 shows some examples of the cubically modelled F-pattern in sets of nine replicates from each recording session. The top speaker shows little variation over one month in his F-pattern. The bottom speaker shows some differences in both F2 and F3.

The cubic polynomial coefficients were processed with the two-level kernel density multivariate likelihood ratio formula described in [11], which was developed to handle dependencies between predictor variables (here, the 12 cubic polynomial coefficients). This formula evaluates the difference between two speech samples against the within- and between-speaker co-variance estimated from a reference sample in order to say whether the difference is more likely assuming same- or different-speaker provenance.

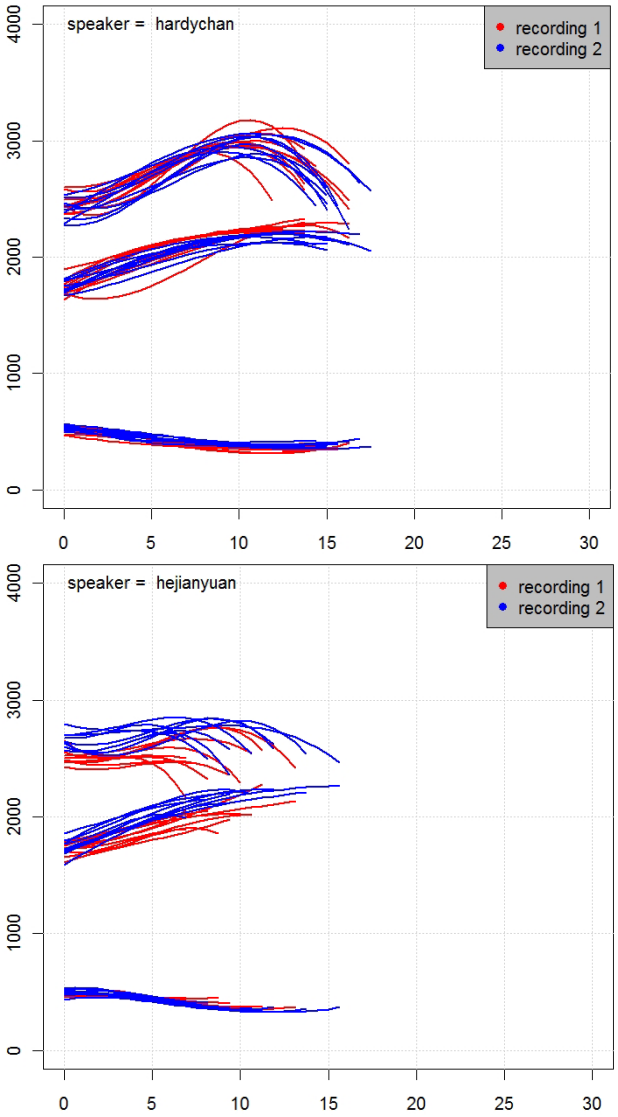

Figure 4: *Examples of non-contemporaneous F-pattern differences (modeled with cubic polynomials) in Cantonese /ei/ for two speakers.*

In calculating LRs, each speaker's data in their first recording was compared both with their data in the second recording to get LRs for same-speaker comparisons, and with the data of other speakers to get LRs for different-speaker comparisons. With two non-contemporaneous recordings, two independent different-speaker LRs are possible for the same two different speakers (e.g. speaker one in recording one vs. speaker two in recording two, and speaker one in recording two vs. speaker two in recording one). We used just one set of different-speaker comparisons, so in all we had 18 LRs from 18 same-speaker comparisons and 153 LRs from different-speaker comparisons. Because of the small number of speakers, we used cross-validation with a leave-one-out method. This means the two samples being compared in order to estimate their likelihood ratio were always removed from the reference sample. If it was a same-speaker comparison, only that speaker was removed; if it was a different-speaker comparison, both speakers were removed. In this way, the reference sample was always independent of the samples being compared. Intrinsic comparisons, where the speakers being compared are not removed from the reference sample, were also carried out to check the effect of cross-validation.

To the extent that the F-pattern trajectories in Cantonese /ei/ are forensically useful, same-speaker comparisons will be evaluated with a LR greater than one (or $\log_{10}LR > 0$); and different-speaker comparisons will be evaluated with a LR less than one (or $\log_{10}LR < 0$).

The set of LRs output from the MVLR formula can be displayed in a Tippett plot, as in Figure 5. This shows the cumulative distribution of same- and different-speaker LRs, with the different speaker LRs ascending to the left and the same-speaker LRs ascending to the right. It can be seen that the discrimination is quite good for a single segment – an equal error rate of ca. 15% – if we were to set the threshold at about $\log_{10}LR = -3$. But as it is, the system is giving us some misleading information: for example, about 10% of same-speaker comparisons were evaluated with a LR between about -5, and -10 which very strongly suggests that the comparison would be exceedingly more likely if the samples had come from different speakers. The high value of 2.41 (well above unity) for the log-likelihood ratio cost (C_{llr}), which is the proper measure of likelihood ratio detection systems [12] reflects this. (C_{llr} values have to be lower than unity to show that the system is giving information.) In order to improve this, the LRs (usually termed 'scores') need to be calibrated with logistic regression, according to [13]. The result is shown in next section.

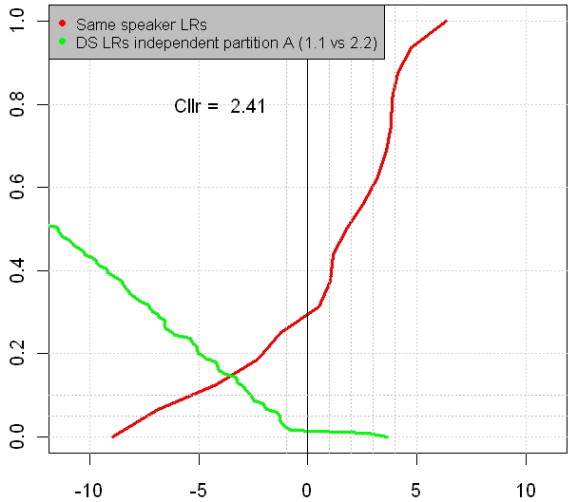

Figure 5: *Tippett plot for uncalibrated cross-validated LRs from Cantonese F-pattern trajectories in /sei/. x axis = $\log_{10}$LR greater than …; y axis = 1- cumulative proportion of same-speaker trials ~ cum. prop. of different speaker trials.*

3. Results

The Tippett plot for the calibrated LRs is shown in Figure 6. Both cross-validated LRs (solid lines) and non-cross-validated LRs (dotted lines) are shown. It can be seen that the same-speaker and different-speaker LRs have been "shifted and scaled" [13], so that the C_{llr} is now much smaller than one – for the cross-validated LRs it is 0.46 – and the same-speaker and different speaker LR curves cross almost exactly on the $\log_{10}$ 0 line (where the equal error rate is now about 14%). As a result, the LRs now do not appear to be as strong as before, but the lower C_{llr} indicates now that the diphthong may carry enough speaker-individualising potential to be of use in forensic voice comparison. The non-cross-validated LRs are very highly correlated with the corresponding cross-validated LRs. The difference in the performance may be due to a reduction in between-speaker variance associated with removing one or two speakers from the reference sample for each comparison.

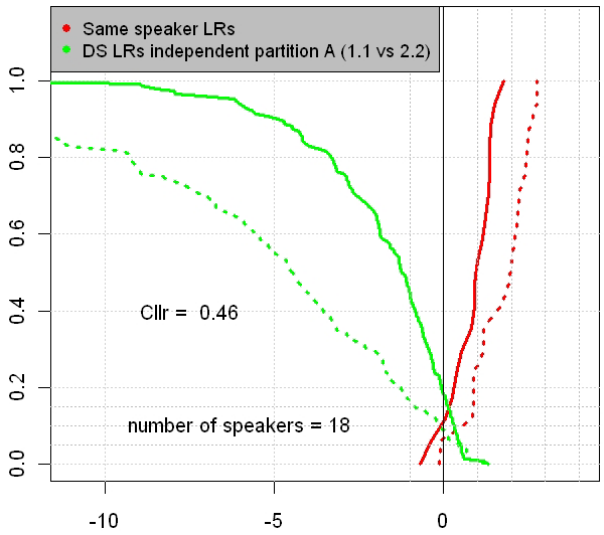

Figure 6: *Tippett plots for calibrated LRs from Cantonese F-pattern trajectories in /sei/. Bottom axis is $\log_{10}$LR greater than .... y axis = 1- cumulative proportion of same-speaker trials ~ cum. prop. of different speaker trials. Dotted lines show LRs from non-cross-validated comparison.*

4. Summary

Following similar experiments on diphthongs in other languages, this paper has described a forensic voice comparison experiment in Cantonese, to investigate the potential of its /ei/ diphthong. Non-contemporaneous recordings of the F-pattern trajectories of /ei/ from 18 male native speakers were analysed with a multivariate LR to see how well the same-speaker and different-speaker comparisons were resolved. A C_{llr} of 0.46 was obtained for the cross-validated calibrated LRs, which shows the diphthongal F-pattern acoustics have forensically useful information, and could possibly be combined with other speech sounds to get better LRs.

The limitations of this study, which make it only a pilot study, are, firstly, the small number of speakers. Many more speakers are needed to get a better idea of the true C_{llr} value for /ei/, since if we had more speakers, there would certainly be more different-speaker pairs who would be similar enough to give LRs bigger than one, and possibly more same-speaker pairs different enough to give LRs less than one. Our recordings are clean, which is unlike forensic reality. More dirty recordings would increase the C_{llr} without doubt. Besides, phone transmission would exclude the lower frequency part of the /ei/ F1. This could be partially handled, however, by just using the F1 intercept.

5. Acknowledgements

We want to thank all the speakers who let their voices be recorded for this paper, and also our friends and colleagues who gave us their database recordings. We also thank the Hong Kong University of Science & Technology for making it possible to run this experiment as part of their Humanities postgraduate course *Topics in Chinese Phonetics: Forensic Voice Comparison in Cantonese*. And very many thanks to our three reviewers for their careful reading, corrections, and useful comments, most of which we have incorporated. This paper was written using findings from *Australian Research Council Discovery Grant* No. DP0774115.

6. References

[1] Morrison, G.S. Forensic voice comparison. In I. Freckelton, & H. Selby [Eds.], Expert Evidence (Ch. 99), Thomson, 2010.

[2] Evett I.W. "Towards a uniform framework for reporting opinions in forensic science casework", Science & Justice, 38(3): 198-202, 1998.

[3] Morrison, G.S. "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories", JASA 125: 2387–2397, 2009.

[4] Rose, P., Kinoshita, Y. & Alderman, T. "Realistic Extrinsic Forensic Speaker Discrimination with the Diphthong /ai/", Proc. 11th Australasian Intl. Conf. on Speech Science & Technology, 329-334, 2006.

[5] Rose, P. "The Intrinsic Forensic Discriminatory Power of Diphthongs", Proceedings of the 11th Australasian International Conference on Speech Science & Technology, 2006.

[6] Zhang, C., Morrison, G.S., & Rose, P. "Forensic speaker recognition of Chinese /i/ and /y/ using likelihood ratios." Proceedings of Interspeech, ISCA:1937–1940, 2008.

[7] Zhang, C., Morrison, G.S., & Thiruvaran, T. **"**Forensic voice comparison using Chinese /iau/." Proc. 17th International Congress of Phonetic Sciences, Hong Kong: 2280–2283, 2011.

[8] Morrison, G. S., Rose, P., and Zhang, C., "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice**"**, Australian Journal of Forensic Sciences, 12: 1-13, 2012.

[9] McDougall, K. "Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies", Intl. Journal Speech Language and the Law, 13(1): 89-126, 2006.

[10] McDougall, K. and Nolan, F. "Discrimination of Speakers using the formant dynamics of /u:/ in British English". In J. Trouvian and W. Barry [Eds.], Proc. 17th Intl. Cong. Phonetic Sciences, Saarbrueken, 1825 – 1828, 2006.

[11] Aitken, C.G.G. & Lucy, D. "Evaluation of trace evidence in the form of multivariate data", Appl. Statistics 53(4): 109-122, 2004.

[12] Morrison, G.S. "Measuring the validity and reliability of forensic likelihood-ratio systems**"**, Science & Justice, 51: 91–98, 2011.

[13] Ramos-Castro, D.; Gonzalez-Rodriguez, J.; Ortega-Garcia, J. "Likelihood Ratio Calibration in a Transparent and Testable Forensic Speaker Recognition Framework", Speaker and Language Recognition Workshop, IEEE Odyssey: 1 – 8, 2006.