Likelihood Ratio-based Forensic Voice Comparison with Cantonese /i/ F-pattern and Tonal F0

Cai Yu Wang[1,2] *& Phil Rose* [1, 3]

[1]Division of Humanities, School of Humanities and Social Science, Hong Kong University of Science and Technology, Hong Kong, China
[2] School of Foreign Languages, Zhongnan University of Economics and Law, Wuhan, China
[3] School of Language Studies, Australian National University
wangcaiyu@ust.hk, philip.rose@anu.edu.au

Abstract

This paper describes a pilot likelihood ratio-based forensic voice comparison experiment using non-contemporaneous samples of Cantonese /i/ from the natural speech of 26 Cantonese males. Two phonetic features of /i/, F-pattern and tonal F0, are used. F-pattern trajectories are parametrically quantified by polynomial coefficients, instead of point values of F1, F2 and F3, and the coefficients used to calculate a two-level kernel density multivariate likelihood ratio. The LR for tonal F0 is calculated in the same way. The log likelihood ratio cost function C_{llr} is applied to evaluate the accuracy of the performance. The individual C_{llr} values from F-pattern and F0 are 0.65 and 0.68 respectively. By fusing the LR of these two features with logistic regression, a small reduction in C_{llr} to 0.59 is achieved. This suggests that both /i/ F-pattern and tonal F0 features might be of use in Cantonese forensic voice comparison.

Index Terms: Forensic voice comparison, likelihood ratio, Cantonese, F-pattern, F0, /i/.

1. Introduction

This research is part of a larger project on the Forensic Voice Comparison (FVC) with Cantonese acoustics. Among the 52 rimes in the Cantonese phonological system [1], seven were chosen in this project covering most of the phonotactic types of Cantonese rime, i.e. monophthong, diphthong, triphthong, syllabic nasal and stopped. Specifically, they are: /i/ [i:], /aai/ [a:i], /eu/ [ɔy], /ei/ [ei], /jat/ [jɐt˚], /jau/ [jɐu] and /m̩ ~ ŋ̩/ [m̩ ~ ŋ̩]. These specific rimes were chosen on the basis of demonstrated [2 5 6] or assumed FVC potential. This paper focuses on the /i/ rime and aims to investigate its FVC potential in Cantonese.

This FVC of Cantonese /i/ is conducted within the logical framework of the likelihood ratio of Bayes' Theorem. The LR is determined by the probability of the evidence under two hypotheses – the prosecution hypothesis that both samples from the offender and suspect are uttered by the same person and the defense hypothesis that the samples are produced by different speakers. The schematic formula of this concept is shown as follows [2]:

$$\frac{p(H_{SS} \mid E_{Sp})}{p(H_{DS} \mid E_{SP})} = \frac{p(H_{SS})}{p(H_{DS})} * \frac{p(E_{Sp} \mid H_{SS})}{p(E_{Sp} \mid H_{DS})}$$

Posterior Odds = *Prior Odds* * *Likelihood Ratio*

According to the *Daubert* Standard, to qualify for admission as scientific evidence, approaches should have been tested and error rates known. The LR approach to FVC has been tested in Spanish, French, Japanese and Australian English and has shown to be effective [3 4]. In Chinese, however, this testing has only started very recently on a few rimes [i, y, jau] of Shenyang Mandarin, a dialect phonetically and phonologically similar to Beijing Mandarin [5 6 7]. Cantonese, one of the phonologically more complicated dialects in China because of its complex rime and tonal systems, and used in almost all the Chinese communities around the world, has not yet been tested. We make a start here.

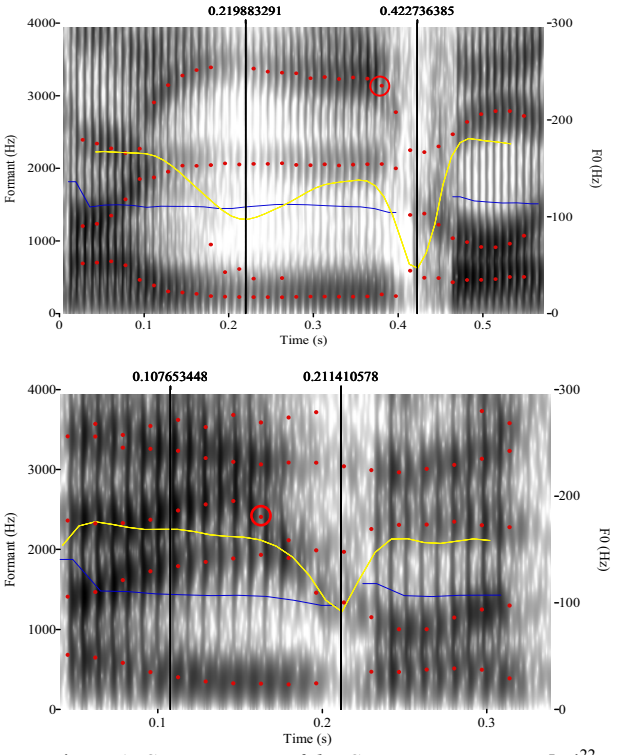

Figure 1: *Spectrograms of the Cantonese utterance* [tai^{22} i^{22} ko^{33}] *from two different speakers (Kahang, Edward) showing between-speaker differences in /i/ F-pattern. red = formants, blue = F0, yellow = intensity.*

The Cantonese /i/, e.g. in the morpheme "二" /ji^{22}/ 'two' is phonologically a monophthong for which one might expect level formant trajectories. But in natural speech, even at a normal speed, it is affected by its linguistic environment and its formants can present a gliding movement as the speaker's supralaryngeal vocal tract moves from and to surrounding vocalic and consonantal targets.

To illustrate this, figure 1 shows the acoustics of /i/ in the utterance [tai^{22} i^{22} ko^{33}] 第二个 *the second one* from two different speakers. For the speaker in the top panel the onset of

the rime is marked clearly by a drop in intensity (this is presumably part of its /j/ onset). As can be seen his /i/ has a fairly stable F-pattern until the 'velar pinching' drop in F3 at the end caused by the following velar stop. The speaker in the bottom panel lacks the intensity drop and has a far more dynamic F-pattern which may be associated with his quicker articulation.

Because of this dynamicity, the conventional way of measuring F-pattern by point values, or by calculating the mean of several point values of the formants (usually F1 and F2), may not be the best way to capture the F-pattern of /i/. As a result, we follow [8] in using F-pattern trajectories instead of point values for calculating LRs, even for this nominal monophthong.

There are nine tones in Cantonese: six in sonorant-final syllables (three level, two rising and one falling); and three (high, mid and low) in syllables ending with unreleased stops [p, t, k]. Is tonal F0 useful in the FVC of tonal languages? Does it show any correlation with other features in FVC? To shed light on these natural questions, we examine the FVC potential of the F-pattern and tonal F0 of the Cantonsese morpheme [i^{22}] 二 *two*.

2. Procedure

For the project on Cantonese FVC, the MTR (Hong Kong Mass Transit Railway) database was built, and the first author of this paper was one of the constructors. This database is designed to get non-contemporaneous controlled natural speech, in which two recordings of each speaker were collected with an interval of 3 to 5 weeks. The experiment described here on Cantonese /i/ uses data from 26 speakers of the MTR database.

2.1. Speakers

The sound data was recorded in Hong Kong from 26 speakers: all male Hong Kong locals. They are either full- or part-time graduates with ages ranging from ca. 23 to 30 except for Speaker 9 who is 45. Besides their mother tongue, Cantonese, most of them are fluent in English and Putonghua (the standard promoted in mainland China). No speakers had a hearing or speech defect.

2.2. Corpus and elicitation

Two lists of 45 questions about the Hong Kong MTR were designed for the two non-contemporaneous recordings to obtain the target rimes. Each list contained two warm-up questions, 40 elicitation questions for the target segments, and three randomly inserted dummy questions.

In each recording, ten different questions were designed to elicit the answer "the second station" ("第二个站" [tai^{22} i^{22} ko^{33} tsaam^{22}]) which contains the target /i/ for this paper. For example: "Is Tin Hau the first or second station after Wan Chai?" Theoretically, provision was made in the task for 10 such answers containing the segment /i/ to be elicited in one recording session. In reality we sometimes got less, sometimes more, due to incorrect answers. In all, however, we obtained between eight and 11 /i/ replicates per speaker per recording.

The recording was done at HKUST in two adjacent rooms with sound-proof walls, one for the researcher and the other for the speaker. The researcher led the conversation over the phone, but the responses of the speaker were recorded directly onto the computer through a high-definition SONY ECM 907 microphone. Mono-channel 41 kHz 16 bit resolution recordings were captured using *Cool Edit Pro* 2.0. To ensure the quality of the recording, another researcher monitored the recording on the computer.

Before the recording, subjects were asked to try to give full answers to the questions by repeating the given information, and generally to say as much as they wanted. Not being a native speaker of Cantonese, the researcher led the conversation in Putonghua, with the speakers responding in Cantonese. As speakers understood Putonghua this bilingual elicitation did not prove a problem. A typical recording session lasted about 12 minutes.

2.3. Within & between-speaker tempo differences

Some between-speaker differences in F-pattern were already shown in figure 1, and as noted there were also differences in tempo. Figure 2 shows some examples. The top panel shows a speaker who was fairly consistent in both F-pattern and tempo across both recordings, with /i/ F-patterns lasting between 10 – 13 csec. The speaker in the bottom panel obviouly spoke much quicker the second time around, and had higher F-pattern values.

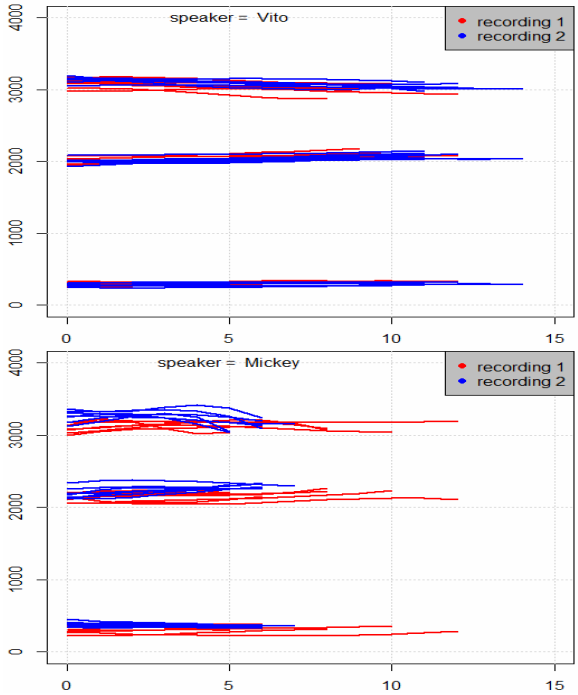

Figure 2: *Between-speaker differences in tempo. /i/ F-pattern trajectories (Hz) plotted as a function of absolute duration (csec) from two speakers in two non-contemporaneous recording sessions (shown in different colours).*

It is not clear how the difference in F-pattern for the second speaker in Fig. 2 relates to the difference in duration, but such variation is of course one of the factors which make FVC difficult. But like other paralinguistic factors which may affect the realization of a phoneme, this cannot be avoided in natural speech or real cases of FVC, and we have to live with them.

All the [tai^{22} i^{22} ko^{33}] utterances (第二个 *the second one*) were identifed and edited out from the 26 speakers' two recordings, and, after visual checking, F1, F2 and F3 centre frequencies over the /i/ segment were extracted with *Praat's formant listing* command. The intensity minimum was used as the prime indicator of onset, as in the top panel of Fig.1. In cases where such minima were not present, especially in quick

speech (as in the bottom panel of Fig 2), the starting point of the extraction was adjudged to be half way between the low first target in /tai/ and the F2 peak. To exclude any effect from the following consonant, the extraction was adjudged to end at the point where F3 starts to decrease for the velar pinch (these points are circled in Fig.1).

Generally we found extracting four formants below four kHz effective. However it was necessary sometimes to adjust the settings to suit the speaker. Any obviously incorrect values were corrected manually (e.g. in the top panel of Fig. 1 the two formant values between the F1 and F2 trajectories at ca. 400 Hz were ignored). Tokens that were not well extracted by *Praat* were discarded. In this way, seven replicates per speaker per recording were obtained. The trajectories of the formant centre-frequencies thus extracted were then modelled by cubic polynomials using code written in *R*.

2.4. Within- & between-speaker tonal F0 differences

The tonal pitch of the Cantonese morpheme [i^{22}] *two* is low level, and it was noted that its F0 shape was generally stable in the natural speech of the corpus (see, e.g. the F0 shapes in Figure 1). To illustrate the between- and within-speaker variation in the 26 speakers' [i^{22}] tonal F0 data, mean F0 values were calculated and plotted in Fig. 3. This figure shows most speakers produce consistent F0 for this morpheme across the non-contemporaneous recordings. But of course many speakers also have similar mean F0 values to others (e.g. speakers 1, 13 and 23). In addition speaker 9 shows an obvious large within-speaker difference.

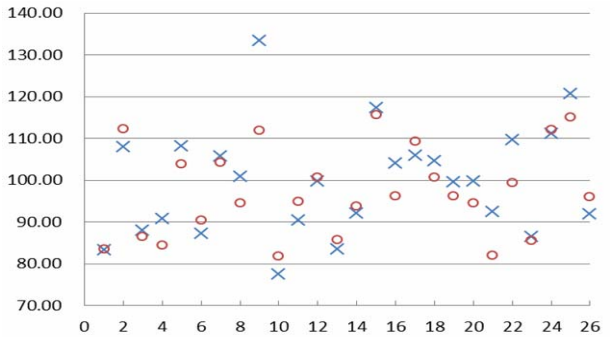

Figure 3: *Between- and within-speaker variation in [i^{22}] mean F0 (y axis) for 26 speakers (x axis). Non-contemporaneous recordings are shown by crosses (rec. 1) and circles (rec.2).*

Analogous to the F-pattern, the F0 trajectories of [i^{22}] were used instead of point estimates. F0 values were extracted by *Praat* with a pitch [sic] range from 50 to 300 Hz. The F0 trajectories were then modelled in *R* with quadratic polynomials (the lower order was based on our assumption that there would be less time-course variation in F0 than in some formants).

2.5. Further processing

The coefficient values from the polynomial modelling of F-pattern and tonal F0 were input into the multivariate likelihood ratio formula from [9] to obtain respectively the F-pattern and tonal F0 LRs. The specific set of coefficients to use was determined empirically using minimum C_{llr} (the accuracy metric for LR-based detection systems [7]). It was found that the best results for F-pattern could be obtained by omitting all F1 information, and also the cubic coefficients of F2 and F3. This is not surprising, given the nature of F-pattern variation in a phonologically monophthongal segment. For tonal F0, the quadratic coefficients could also be discarded: again, the F0 time-course corresponding to the tonologically level pitch does not apparently require such detailed modelling. Cross-validation was performed with a leave-one-out procedure, but we have also given corresponding non-cross-calibrated results for comparison. The raw LRs were then calibrated with logistic regression. Logistic-regression fusion [10] was then used to combine the LRs from F-pattern and F0.

3. Results

As expected, the raw LRs showed good discrimination but bad calibration, with the C_{llr} values considerably reduced post-calibration. The calibrated LRs for F-pattern and tonal F0 are presented separately in Tippett plots below which enable us to see to what extent same-speaker comparisons produce LRs greater than $\log_{10}$ 0, and *vice-versa*; and to inspect the magnitude of the counterfactual LRs.

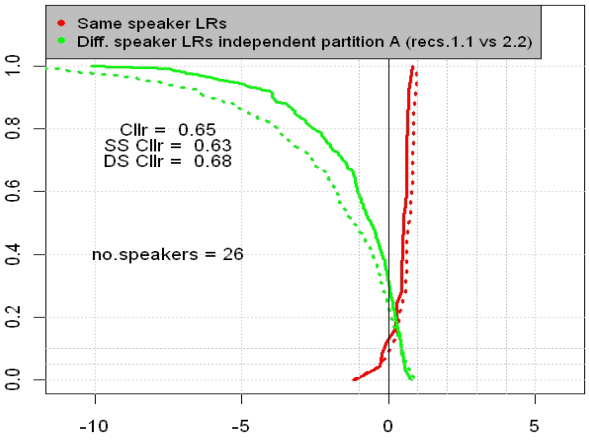

Figure 4: *Tippett plot for Cantonese /i/ F-pattern. x axis = log_{10}LR greater than …; y axis = 1- cumulative proportion of same-speaker trials ~ cum. prop. of different speaker trials; dotted lines = non-cross-validated, solid lines = cross-validated comparisons.*

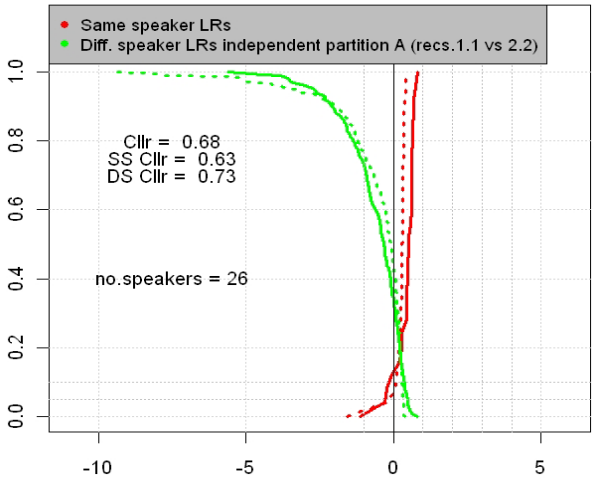

Figure 5: *Tippet plot for Cantonese /i/ tonal F0. Information as for figure 4.*

Fig. 4 shows the results for the F-pattern (actually just F2 and F3, since, as explained, F1 coefficents were not used). It

can be seen that the more extreme cross-validated LRs (solid lines) are bracketed by the non-cross-validated LRs (dotted lines), and so failure to cross-validate would lead to a slight over-estimate of the performance. We can see, with a $\log_{10}$LR greater than the $\log_{10}$LR=0 threshold, ca. 12% of the same-speaker LRs are evaluated as more likely had they come from different speakers, and about 30% of different-speaker LRs are counter-factually evaluated. The magnitude of these counter-factual LRs is not particularly great, however, and this, together with the amount of incorrect evaluations, gives a useful, but certainly not fantasitc, C_{llr} for /i/ F-pattern of about 0.65. Treating this as a discrimination, the equal error rate (EER) is about 18%. The magnitude of strength of evidence for same-speaker comparisons does not exceed $\log_{10}$LR 1, and is very weak.

Fig. 5 shows the results for tonal F0. Interestingly, the cross-validated results are slightly better than the non-cross-validated. About 14% same-speaker LRs and some 34% different-speaker LRs are against the ground truth. The performance is a little worse than the F-pattern, with a C_{llr} of 0.68, and an EER of ca. 19%, and weak same-speaker strength of evidence.

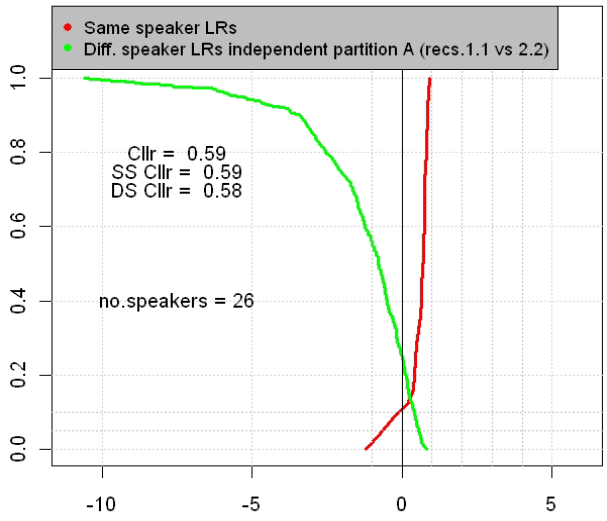

Figure 6: *Tippett plot for cross-validated fused Cantonese /i/ F-pattern and tonal F0.*

As shown in Fig. 6, fusing the /i/ F-pattern and tonal F0 results in a slight improvement, indicating that there is therefore a small amount of speaker-specific information not shared between tonal F0 and F-pattern in /i/. The C_{llr} value is reduced a little to 0.59, with ca. 11% LR counterfactual same-speaker comparisons, and ca. 25% counterfactual different-speaker comparisons, and an EER of ca. 12%. Strength of evidence for same-speaker comparisons remains weak.

3.1. Summary

The forensic voice comparison experiment with Cantonese /i/ described above has shown that both F-pattern and tonal F0 trajectories yield individual C_{llr} values less than unity which also reduce a little upon fusion. They may therefore be of some use in forensic voice comparison in Cantonese.

This study has many limitations. First of all, the recordings were clean, and not contaminated by telephone, especially mobile phone transmission. With phone recordings, information from /i/'s low F1 would be compromised. However, since we discarded F1 coefficients the clean nature of the recordings will not have had much of an influence in this respect. Mobile phone degradation always remains a worry, however, and it would be interesting to see what happens if the data were processed through a mobile phone network first: perhaps the F0 would become relatively more important under such circumstances.

More important is the fact that, although the /i/ tokens were taken from recordings of natural, unedited speech, they were nevertheless tightly controlled by occurring in a fixed phonological environment. This undoubtedly contributed to the performance of the F0, and we suspect that if we relaxed the conditions, its contribution might be lessened. As far as the F-pattern is concerned, if we took [i] tokens from a wider set of environments, the F-pattern trajectories might also be compromised, and it would be interesting to see how much information is lost by reverting to the old point measurement on the F-pattern, rather than a trajectory.

In sum, the results from just this single segment, although not stellar, are at least encouraging, and /i/ could probably contribute a little to an overall LR, but only in conjunction with other segments.

4. Acknowledgements

The first author sincerely thanks the graduates of Humanities in the Hong Kong University of Science and Technology for their warm help, both the volunteer speakers and those who vacated their office for the recordings. We also thank the Hong Kong University of Science and Technology for making it possible to run this experiment as part of their postgraduate Humanities course *Topics in Chinese Phonetics: Forensic Voice Comparison in Cantonese*. Thank you also to our reviewers for taking their time to make very useful comments, most of which we have incorporated. This paper was written using findings from *Australian Research Council Discovery Grant* No. DP0774115.

5. References

[1] Cheung, H. S., "Cantonese of the 21st Century: formation of a New Phonological System in the Hong Kong Language", Journal of Jinan University (Philosophy & Social Science Edition), 24(2), 25-40, 2002.

[2] Rose, P., Forensic speaker discrimination with Australian English vowel acoustics, Proceedings of the 16th International Congress on Phonetic Sciences, Saarbrucken, 1817-1820, 2007.

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

[4] Zhang, C. L. and Rose, P., "Strength evaluation of forensic speaker recognition evidence based on Likelihood Ratio approach", Evidence Science, 16(3), 337-342, 2008.

[5] 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.

[6] 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.

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

[8] Morrison, G. S., Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs, Acoustical Society of America, 2387-2397, 2009.

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

[10] Brümmer, N. et al., "Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006," IEEE Trans. Audio, Speech, Lang. Process. 15, 2072-2084, 2007.