

Likelihood Ratio-based Forensic Voice Comparison with the Cantonese Triphthong /iau/

Aishu Chen^{1,2} Phil Rose^{2,3}

¹Department of Linguistics and Modern Languages, Chinese University of Hong Kong

²Department of Humanities, Hong Kong University of Science & Technology

³School of Languages, Australian National University

aishu.chen@gmail.com, philip.rose@anu.edu.au

Abstract

This paper describes a likelihood ratio-based forensic voice comparison using the Cantonese triphthong /iau/. The features for comparison are both the formant pattern and tonal F0 of /iau/ contained in 15 native speakers' non-contemporaneous natural speech. Kernel density multivariate likelihood ratios are calculated based on polynomial coefficients converted from raw frequency values of F1, F2, F3 and F0. A log-likelihood-ratio cost (C_{lr}) value for the F-pattern of 0.6 is obtained, showing that comparisons with /iau/ F-pattern have reasonable validity, and also reasonable strength of evidence can be achieved. Likelihood ratios calculated based on tonal F0 do not show good performance ($C_{lr}=0.95$), and are therefore not fused with the F-pattern LRs.

Index Terms: Forensic Voice Comparison, Likelihood Ratio, Cantonese, /iau/, tonal F0, F-pattern

1. Introduction

The Forensic Voice Comparison (FVC) described in this paper is done with a likelihood ratio-based model on the Cantonese triphthong /iau/ (phonotactically /GVG/). Assuming a FVC scenario, a recording containing the target /iau/ from an unknown speaker is to be compared with recordings of a known voice which also include /iau/. The unknown speaker is often the offender, while the known voice is usually the suspect. What the trier of fact is interested in is whether the unknown voice belongs to the same speaker as the known voice. The two voice samples could be attempted to be discriminated by a naïve listener (perhaps a policeman), but the scientific way is to compare the voice technically, using acoustic-phonetic features extracted from the samples. However, even scientific comparisons may not give the jury definite answers, because speakers' voices are so complex. In speech acoustics, for example continuous between-speaker and within-speaker variation obtains. To illustrate this, some between- and within-speaker differences in /iau/ are shown in figures 1 and 2.

The two spectrograms in figure 1 are from the same speaker CKW's non-contemporaneous recordings. The syllable /iau/ was taken from the beginning of a disyllabic word for a Hong Kong place name: *Yau Tong*. It can be seen that the F-patterns of these two tokens of /iau/ are not the same. Especially, the two trajectories of F3 appear to differ. At the lowest point of F2, the frequency difference between F2 and F3 is considerably bigger in the second spectrogram of /iau/. Also in the second spectrogram a zero at about 2400 Hz cuts through the falling part of F3. These variations indicate slightly different tongue movements of the same speaker, and also possible different interaction with the sub-glottal system

related to the low tonal F0 of this word, when producing these two tokens.

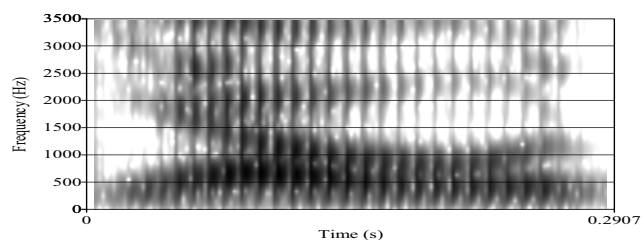
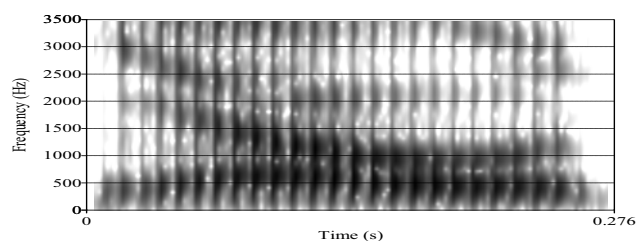


Figure 1: Spectrograms of /iau/ from non-contemporaneous recordings of a single speaker (CKW).

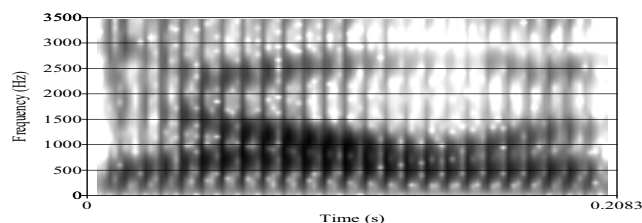


Figure 2: Spectrogram of /iau/ from speaker LJH.

The F-pattern of /iau/ from another speaker, in figure 2, is different from that of CKW. LJH's F3 distributes in a higher frequency range (above 2500 Hz), and the turning point of F2 and F3 starts earlier, indicating earlier anticipation of the alveolar place of the following consonant. However, otherwise F1 and F2 trajectories seem to be quite similar as CKW's first two formants' trajectories.

How do we make sense of these differences and similarities so that it is possible to say that the first two come from the same speaker and the third is from a different speaker? The logically correct conceptual framework for this is Bayes' Theorem, which says that the posterior probability in favour of a hypothesis is proportional to the strength of the evidence adduced in its support and the prior probability of the hypothesis before the evidence is adduced. Since the forensic voice comparison expert does not normally know the prior probability, their job is to use the Likelihood Ratio (LR) to estimate the strength of evidence involved [4, 5, 7, 8]. If the

prior odds are known, it is possible to use Bayes' Theorem to say how likely it is that the two voice samples are from the same speaker. But this is a decision for the trier of fact, not the expert, and this partitioning also makes the use of a LR legally correct. Using LRs also contributes to satisfying the requirements of the *Daubert* ruling [13], which says that the experts' methods leading to their testimonies should have been tested. An example of how LRs can be used in such testing is given in this paper.

Two important components are involved in the LR-based FVC model. They are the evidence and the competing hypotheses. These can be seen in the formula at (1) for calculating a LR (adapted from [7, 9]):

$$LR = \frac{p(E_{sp}|H_{SS})}{p(E_{sp}|H_{DS})} \quad (1)$$

The LR can be seen to be equal to the ratio of the probabilities of the difference between two voice samples – the speech evidence E_{sp} – under the same-speaker and different-speaker hypotheses (H_{SS} H_{DS}). Suppose we need to compare two sets of samples of /iau/ based on their F-pattern, a forensic comparison of /iau/ should give us a LR value indicating whether the difference in the F-pattern between the two sets of /iau/ is more likely assuming they come from the same speaker or from different speakers. A LR value bigger than 1 indicates the evidence supports a same-speaker source for the two sets. Conversely, a LR smaller than 1 says that the evidence is in favour of a different-speaker source. The strength of the evidence supporting same-speaker or different-speaker hypotheses is reflected in the magnitude of LR. Suppose in a real criminal case, 1 out of 6 people including the suspect may be the offender. In the absence of any other information or assumption, the prior odds are then 1:5 against the suspect being the offender. If a LR value of 100 is obtained for the evidence, the posterior odds is the prior odds multiplied by the LR [7, 9], which gives (1:5 * 100:1 =) 20:1. When the odds are converted into probability, the posterior probability that the suspect and the offender are the same person would be (20/21 =) ca. 95%.

In this paper we investigate the range in strength of evidence obtained from acoustic-phonetic features of Cantonese /iau/. This triphthong is chosen because a previous study on /iau/ in Shenyang Mandarin [5] has demonstrated that diphthongs have considerable forensic discriminatory power, presumably because they have more information than diphthongs or monophthongs. (It is not clear whether Cantonese /iau/ is phonetically the same as Mandarin /iau/, although from casual observation of formant trajectories, they seem to be different). Apart from the investigation on Shenyang /iau/, only two other Chinese vowels – [i] and [y], again from Shenyang Mandarin – have been tested [4], and so nothing as yet has been done on Cantonese (or any other Chinese variety). This paper thus makes a start, together with other papers in these proceedings e.g. [11], in doing LR-based FVC in Cantonese.

2. Procedure

2.1. Speakers

Our recordings were collected from 15 young male Hong Kong Cantonese native speakers, aged from 18 to 30, all students at the Chinese University of Hong Kong (CUHK) or

Hong Kong University of Science and Technology (UST). Although the subject pool is not large, and therefore requires cross-validated testing, the subjects do have homogenous linguistic backgrounds and speech.

2.2. Corpus

The Cantonese MTR (Mass Transit Railway) database was used to elicit speakers' natural speech. This database was obtained using quasi map-task questions such as "How many stations are there between *Yau Tong* and *Kwun Tong*?" or "How do I get to *Tsim Sha Tsui* from *Yau Tong*?" These questions elicited spontaneous answers from subjects when they were referring to the MTR map, and was a controlled way of getting the target tokens from natural speech. As is essential with FVC testing, two non-contemporaneous recordings were made. Questions in the second recording session were different from the first in order to prevent the speakers from getting familiar with the tasks. Also, a fairly long break of four weeks was set between two recording sessions in the expectation that the break was long enough to introduce a realistic amount of within-speaker variation associated with non-contemporaneous recordings, and also to avoid familiarisation effects. Speakers' answers containing target /iau/ were edited out for further analysis.

Our target triphthongs occurred in station names *Yau Tong* 油塘 and *Yau Ma Tei*. 油麻地, both of which have the same morpheme (or at least same character) with the same low-falling or low level tone: 油/iau 21 ~ 11/. Six tokens in the utterance-initial positions were chosen for F-pattern measurements: three in *Yau Tong*, and three in *Yau Ma Tei*. The factors which may affect the F-pattern of our target /iau/ are coarticulation with the following alveolar or bilabial place in /t/ and /m/, and nasality in /m/.

2.3. Elicitation

Speakers were recorded at CUHK and UST. The CUHK recordings were conducted in a double-walled, sound-attenuated chamber (IAC, type 1200A) located in the Child Language Acquisition Lab, or a soundproof research lab in the CUHK Linguistics department. Recordings were made using a solid state Zoom H2 handy recorder (with a microphone in it). A recording software in an iMac was also at work during the recording to create a backup. Speech was recorded with a sampling rate of 44100Hz. subjects' answers were saved in Windows PCM .wav format on hard disks. At UST, recordings were made with a high quality lapel mike feeding into a Edirol UA 25EX digitiser in a quiet, acoustically absorbent room.

Speakers were seated in front of a desk with the MTR map and the recorder on the desk. We tried to assign speakers into pairs so they could conduct Q&A between them. If they did not come in pairs, the experimenter asked the questions. Speakers were required to sit separately with some space, and no eye contact, between them. They were instructed to pretend talking to each other over the phone. There was a training session before the real recording session where subjects were trained to give full answers to the questions, repeating the station names together with numbers in their answers. Since some speakers included unwanted stations in their counting, they were also trained to count the station numbers in our specified way (excluding the beginning and ending stations). After the speakers understood the tasks, we initiated the H2 handy recorder.

Speakers generally performed well in these tasks. Sometimes they forgot to give full answers, in which case the

experimenter would remind them and repeat the question-answer again. Because the task requires participants to count the numbers of stations, they needed a certain amount of response time to locate the station name on the MTR map and to count numbers of stations. This resulted in some interesting pause phenomena, speakers often prolonging the final /u/. Because of the obvious affect on the token's F-pattern such tokens were excluded. It took less than 20 minutes for the native speakers to finish their two rounds of Q&A. After the recording, most speakers reported they were able to manage the tasks. Some of them were interested in the purposes of the study, and they were told this after their second recording.

2.4. Measurement

The recordings were first manually edited in PRAAT [2], the experimenter locating utterances containing target triphthongs, and saving them as separate sound files. Next, target /iau/ tokens were identified by examining the F-pattern of the spectrogram using PRAAT. The starting point of /iau/ was set at the position where clear F2 and F3 trajectories began to emerge. The ending point was marked at F2 minimum, so as to maximally exclude any F2 offset perturbations due to differing place of following consonant (no attempt was made to handle differences due to acoustic nasalisation). Then formant values from F1, F2 and F3 were extracted using PRAAT's *formant listing* function. The default formant setting was to track four formants under 4000 Hz. However, for some speakers the setting needed to be changed to achieve better tracking (as judged from visual comparison with the spectrogram formants). The extracted formant values were examined by the experimenter, and obviously incorrect ones manually corrected.

Besides formant values, we also extracted tonal F0 values from the same tokens of /iau/ used for F-pattern measurements. Only the F0 stable part of each rhyme was selected. The F0 values and time points were extracted using PRAAT. Obviously incorrect values were corrected manually.

The raw formant values were converted to four cubic polynomial coefficients (intercept, t , t^2 , t^3) using *R* code, giving twelve parameters in total. LR was calculated using these cubic polynomial coefficients. We chose to use parametric curves because [5], using Australian English diphthongs, had shown that using parametric curves to capture formant trajectories instead of point measurements of the assumed raw formant target values would improve system performance.

LRs were calculated using the multivariate-kernel density (MVKD) formula in [1], which was expected to handle within-segment correlation. The MVKD formula evaluates the similarity of two voice samples based on the within-speaker variation estimated from a reference population. It also evaluates the typicality of the difference between the voice samples by referring to the between-speaker probability density of the same reference population. Within-speaker (co)variation is estimated with a normal distribution, while between-speaker (co)variation is estimated with a kernel model.

Because of the small number of speakers involved, leave-one-out cross-validation was used. We have also given the non-cross validated results: as they show much better performance the comparison is instructive. The next step of the study would be to increase the number of speakers, so as to be able to use separate data for reference and test material.

Comparisons were made between each speaker's first recording /iau/ tokens and their /iau/ tokens in the second

recording, giving 15 same-speaker comparison LRs. Also, each speaker's first recording /iau/ tokens were compared with every other speaker's second recording /iau/ tokens, giving 105 different-speaker comparison LR estimates.

Calibration of the raw same-speaker and different-speaker LRs was then done in MATLAB by shifting and scaling them with logistic regression so that the best log likelihood ratio cost C_{llr} is achieved [9]. (C_{llr} is a measure of validity used to evaluate likelihood ratio-based detection systems [12].) The calibrated \log_{10} LRs were then plotted in a Tippett plot.

The LRs based on the parametric curve of the tonal F0 were processed in the same way as the F-pattern, except that quadratic instead of cubic polynomials were used.

3. Results

3.1. /iau/ F-pattern

Figure 3 shows the Tippett plot for the cross-validated /iau/ F-pattern. It can be seen that about 18% of same-speaker comparisons were evaluated with a LR that would be more likely assuming different speakers were involved; and ca. 27% of different-speaker comparisons were counter-factually evaluated. There is rather poor strength of evidence for same-speaker comparisons: the maximum LR is only a little bigger than 10. As might be expected from these values, the C_{llr} value is quite high, at 0.6. Although, being less than unity, this indicates that the system is giving information, it is not as low as might be hoped from the F-pattern acoustics of a triphthong. It is interesting to note the large difference between the cross-validated and non-cross-validated (dotted line) results, which are very strongly correlated. These had a much smaller C_{llr} of 0.15. The difference may be due to a reduction in the between-to within-speaker variance ratio of the multivariate likelihood ratio formula caused by the of necessity smaller reference sample used in cross-validation.

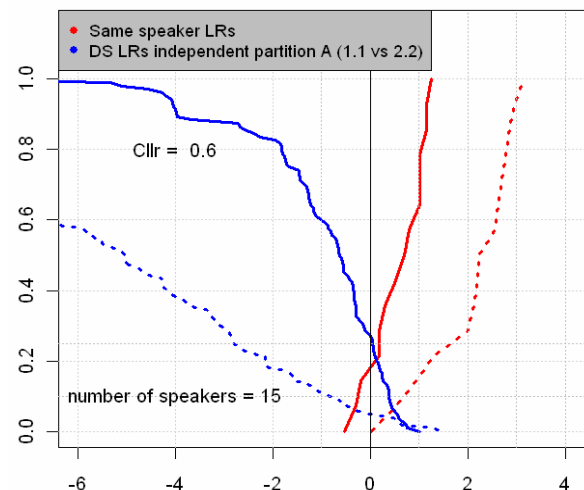


Figure 3: Tippett plot for calibrated LRs from Cantonese /iau/ F-pattern. x axis = $\log_{10}LR$ greater than ...; y axis = 1- cumulative proportion of same-speaker trials ~ cum. prop. of different speaker trials; dotted lines = non-cross-validated, solid lines = cross-validated comparisons.

3.2. /iau/ tonal F0

Regarding the results of FVC using tonal F0, they are very poor, with a calibrated C_{llr} value of 0.95, showing that the

tonal F0 gives us almost no information. The calibrated cross-validated LRs based on tonal F0 comparisons are plotted in figure 4. We originally intended to combine the LRs of F-pattern comparison with LRs of tonal F0 comparison. It is possible to fuse two sets of LRs through logistic regression by taking the correlation between the F-pattern LRs and tonal F0 LRs into account [6]. Given this poor result, however, we did not try to combine the two features based on F-pattern and F0, as we thought it would either give a minimal improvement, or result in worse calibration.

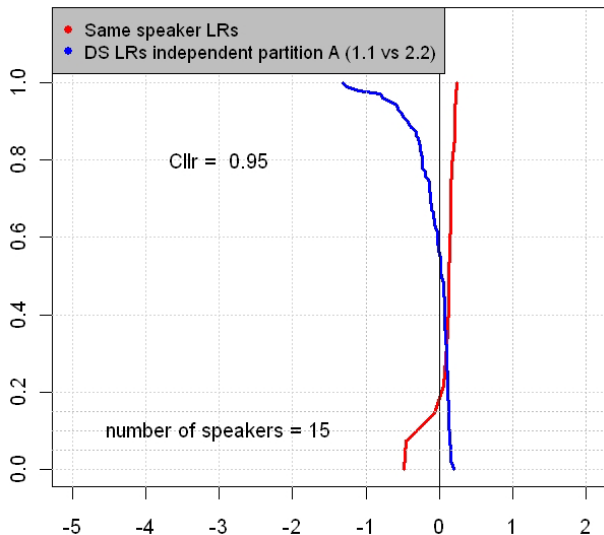


Figure 4: Tippett plot for calibrated LRs for Cantonese /iau/ tonal F0. x axis = $\log_{10}LR$ greater than ...; y axis = 1-cumulative proportion of same-speaker trials \sim cum. prop. of different speaker trials.

4. Summary

The aim of this pilot experiment was to examine whether features like F-pattern and tonal F0 in Cantonese /iau/ are of use in likelihood ratio-based forensic voice comparison. Based on the results obtained, we can conclude that the F-pattern of triphthong /iau/, with a C_{llr} value of 0.6, would probably be useful. However, the C_{llr} value of 0.95 for the tonal F0 comparison indicates that the F0 evidence is not worth using.

The results of this study suggest that in real forensic cases involving Cantonese, using the F-pattern of the triphthong /iau/ as evidence could give useful information to the trier of fact when combined with LRs from other segments. For an example of such a combination of two segments [11] contains the result of combining the F-pattern in /iau/ from this study and the F-pattern of a Cantonese diphthong /oy/, where a C_{llr} of 0.44 is obtained. This demonstrates the usefulness of the F-pattern of /iau/.

The small number of speakers we used must be counted as the major limitation of the current study. Another limitation lies in the well-controlled elicitation conditions, in particular the controlled prosodic and segmental environment. Also we have used clean recordings: over the phone the quality of the sound will degrade, and also we would not be able to use information from most of F1, since it corresponds to high vocalic segments [j] and [w]. All of these factors will increase the C_{llr} . It would be necessary to put these recordings through a phone network and retest. Finally, we have not used discrete cosine coefficients for modelling the F-pattern trajectories, and it may be that they are better than polynomial coefficients for /iau/.

5. Acknowledgements

This paper was written using findings from *Australian Research Council Discovery Grant* No. DP0774115. Thanks firstly should go to all those speakers who participated in the MTR database collection. We also thank the Department of Linguistics and Modern Languages at CUHK for providing a recording booth for us, and the UST for making it possible to run this experiment as part of their Humanities postgraduate course *Topics in Chinese Phonetics: Forensic Voice Comparison in Cantonese*. We also thank our reviewers for taking the time to make very useful comments, many of which we tried to incorporate in our revised paper.

6. References

- [1] Aitken, C. G. G., and Lucy, D. "Evaluation of trace evidence in the form of multivariate data," *Appl. Stat.* 54: 109-122, 2004.
- [2] Boersma, P. "PRAAT, a system for doing phonetics by computer", *Glott International*, 5: 341-345, 2001.
- [3] Brümmer, N., and du Preez, J. "Application independent evaluation of speaker detection," *Computer Speech and Language*, 20: 230-275, 2006.
- [4] 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.
- [5] Morrison, G.S. 2009. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *J. Acoust. Soc. Americ.* 125: 2387-2397.
- [6] Pigeon, S., Druyts, P., Verlinde., "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions", *Digital Signal Processing*, (10) 1-3: 237-248, 2000.
- [7] Morrison, G.S. Forensic voice comparison. In I. Freckelton, & H. Selby [Eds.], *Expert Evidence* (Ch. 99), Thomson, 2010.
- [8] Rose, P. and Morrison, G.S. "A response to the UK position statement on forensic speaker comparison", *International Journal of Speech, Language and the Law*, 16: 139-163, 2009.
- [9] 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.
- [10] Zhang, C., Morrison, G. S., & Thiruvaran, T. "Forensic voice comparison using Chinese /iau/", in W.-S. Lee & E. Zee [Eds.], *Proceedings of the 17th International Congress of Phonetic Sciences*: 2280-228, 2011.
- [11] Li, J. and Rose, P. "Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /oy/ Diphthong", *13th Australasian International Conference on Speech Science and Technology*, Sydney, 2012.
- [12] Morrison, G.S. "Measuring the validity and reliability of forensic likelihood-ratio systems", *Science & Justice*, 51: 91-98, 2011.
- [13] *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993) 113 S Ct 2786. 1993.