Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /ɔy/ Diphthong

Jingwen Li[1,2], *Phil Rose*[2,3]

[1] Department of Linguistics and Modern Languages, Chinese University of Hong Kong
[2]Department of Humanities, Hong Kong University of Science & Technology
[3]School of Languages, Australian National University
joanneljw@gmail.com, philip.rose@anu.edu.au

Abstract

This paper describes a likelihood ratio-based forensic voice comparison experiment using non-contemporaneous natural speech samples of the Cantonese diphthong /ɔy/ elicited from 15 young male Cantonese speakers. Kernel density multivariate likelihood ratios used on polynomial coefficients of the F-pattern trajectories of /ɔy/ give a log-likelihood-ratio cost of 0.55, with a calibrated equal error rate of about 20%. The associated tonal fundamental frequency is also tested but found to be extremely poor. The potential of /ɔy/ is demonstrated by fusion with the Cantonese triphthong /iau/ to get a lower log-likelihood-ratio cost of 0.44 with equal error rate of about 12%.

Index Terms: Forensic voice comparison, likelihood ratio, Cantonese, diphthong, triphthong, F-pattern, fundamental frequency, fusion.

1. Introduction

This paper describes some likelihood ratio-based forensic voice comparison experiments, done in order to investigate the potential of diphthongal F-pattern and tonal fundamental frequency (F0) in Cantonese natural speech.

Forensic voice comparison is different from individual identification or verification in that the latter assume a posterior probability, using which a binary verification or identification can be made. However, in forensic voice comparison the prior probability is not usually known by the forensic expert. Therefore, assuming Bayes' theorem, ("Posterior Probability is proportional to Prior Probability * Likelihood Ratio"), they cannot give posterior probability. They can estimate the Likelihood Ratio (LR), however. The LR is the ratio of the probabilities of the evidence under competing prosecution and defence hypothesis – it quantifies the strength of the evidence in support of the hypothesis that the speech samples come from the same speaker or the hypothesis that they are from different speakers [7].

There have been many LR-based forensic voice comparison experiments, mostly on English and Japanese, using LRs to test the usefulness of various acoustic speech features, e.g. [1, 2]. These have shown that speech acoustics are useful forensically. Due to the *Daubert* requirements on admissibility of forensic scientific evidence [9] stating that approaches have to have been tested before they can be admitted, such testing is rather important.

Similar testing on Chinese is rare: so far there have only been studies on Shenyang Mandarin monophthongs /y/, /i/ [3, 4] and the triphthong /iau/ [5]. Testing on Cantonese, another Chinese Language, has not been done before. Therefore based on the previous research we are now choosing non-contemporaneous speech samples of the Cantonese diphthong /ɔy/, expecting from previous work on diphthongs that the result will be useful because there is potentially more information in diphthongal acoustics. Together with the F-pattern of /ɔy/, the current study is also going to see if tonal F0 is of use. In addition, to demonstrate the usefulness of the diphthong /ɔy/, the study also combines it with a triphthong /iau/ to see how well they perform together.

2. Procedure

2.1. Speakers

As part of a larger project on Cantonese forensic speech comparison, the data was collected from 15 young male native speakers of Hong Kong Cantonese aged between 18 and 30. All the subjects are undergraduate or postgraduate students at the Chinese University of Hong Kong (CUHK) or the Hong Kong University of Science and Technology (UST). It has been controlled during the recruitment that all the subjects were born and brought up in Hong Kong, with Hong Kong Cantonese as their first language. Therefore, it is reasonable to believe that the variety is homogeneous with respect to the features we are quantifying.

2.2. Corpus

Questions about the Hong Kong Mass Transit Railway (MTR) stations were used for the elicitation, and they were of three types: 1. *How many stations are there between A and B?* 2. *Is A the first or the second station after B?* 3. *How do you get from A to B?* Some of the MTR station names were designed to include /ɔy/, the target feature of the present study, and it is expected to be elicited from the answers of the subjects contained in the station names such as *Shum Shui Po* /sam^{1} sɔy^{2} pɔu^{4}/, *Tai Shui Hang* /taai^{6} sɔy^{2} haaŋ^{1}/ and *Sheung Shui* / sɔyŋ^{4} sɔy^{2}/. The /sɔy^{2}/ in these names is the morpheme 水{water}, and it can be seen that it occurs in the middle or end of the word. As it is the same morpheme, it carries the same low-to-high rising toneme, and we used this also to test its tonal fundamental frequency, which we don't think has yet been done for natural speech in a tone language.

The samples were controlled natural speech recorded at two separate sessions. There was a one-month time interval between the two recordings, and there are two reasons for doing this: firstly, in real forensic voice comparison cases the speech samples from the offender and those from the suspect are naturally obtained at separate times; secondly, the estimation of the strength of the evidence may be exaggerated in speech samples from the same recording session, since within-speaker variation is expected to be less in contemporaneous speech samples than in non-contemporaneous speech samples. It was assumed that a one-

month delay was long enough to offset any familiarization effect and avoid the possibility that the subjects answered more fluently at the second recording. However, to be sure, changes were also made to the test materials for the second recording, keeping the target features (station names) but changing the way the information was elicited. The elicitation questions were designed to be in the form of questions and answers so that the subjects would focus on carrying-out the task such as counting how many stations there are between station A and station B, rather than focusing on the fact that they were doing an experiment. Therefore it is believed that the speech is sufficiently natural, though its content is controlled in the areas that we are interested in. All the data in the present study is included in the MTR database.

2.3. Elicitation

A map of the Hong Kong MTR with Chinese station names was put on an iPad to show to the subjects before the actual recording. Two paired-up participants, with one asking and the other responding, elicited the spontaneous speech samples, and then they were swapped around. To be more realistic, the two participants were seated apart with no eye contact, pretending to be having a telephone conversation. In a few cases it was not possible to pair-up subjects, and then the experimenters (who are not native Cantonese speakers) did the elicitation in Mandarin or English, after having practiced with the participants beforehand to make sure they could clearly understand the questions in Mandarin/English and response naturally in Cantonese. Given the typical high fluency of the subjects in Mandarin and English this turned out not to present any problems.

Speakers were recorded at the CUHK or the UST. 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 of the CUHK Linguistics department. Recordings were made using the built-in microphone of a Zoom H2 solid-state recorder, and saved to hard disc as 44.1 kHz 16 bit .wav files. The recorder was placed 20 cm to 30 cm in front of the speaker. In addition, a MacBook Pro computer was also at work during the recording to create a backup. At UST, recordings were made with a high quality lapel mike feeding into an Edirol UA25EX digitiser in a quiet, acoustically absorbent room.

The speakers were instructed before the real recording that they should speak normally and try to give a complete answer (i.e. including the station names) to the questions. They could get familiar with the stations on the map in advance for as long as they wanted.

Most speakers performed satisfyingly with regard to the production of natural speech, and answering with complete answers. If they did not, they were prompted to give a complete answer. This enabled us to get between 7 and 10 useable tokens of /sɔy^2/ for analysis from each of the 15 speakers at each recording.

2.4. Measurement

The diphthong /ɔy/ was chosen for the investigation because as a diphthong with a low back rounded first target and a high front rounded second target, its F-pattern includes substantial movement, and there is some acoustic space for speakers to differ in. Although its onset may contain subglottal effects from the voiceless initial consonant /s/, as a mostly high vowel we do not expect any intrinsic nasalization, and generally we hope that its F-pattern is easily extracted. Figure 1 shows two wideband spectrograms of /sɔy^2/ from two speakers Jeffery and Ng, in order to illustrate typical F-pattern acoustics and between-speaker differences. Both tokens occurred in the middle of the station name *Shum Shui Po*, so the /ɔy^2/ was followed by a bilabial stop.

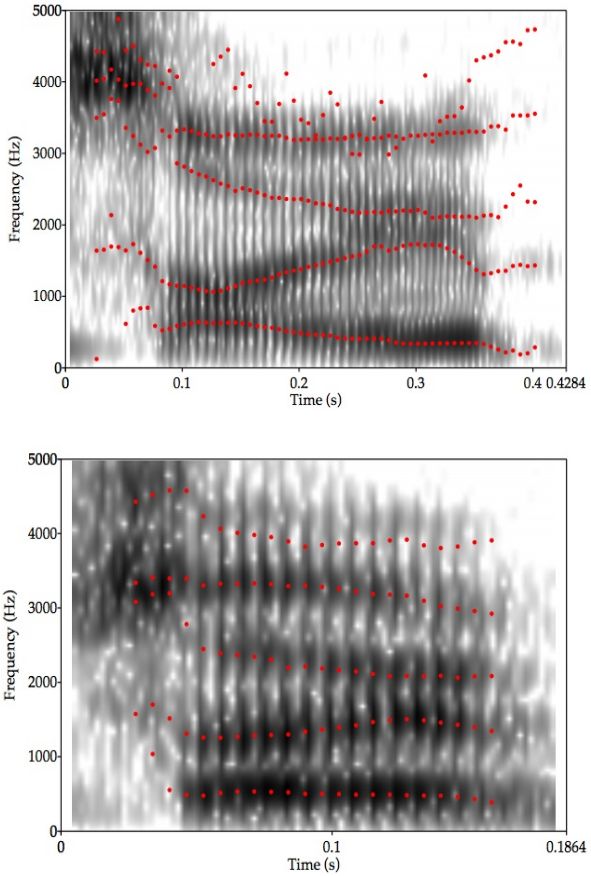

Figure 1: *Spectrograms, with superimposed formant centre-frequencies, of /ɔy^2/ from two different speakers. Top is Jeffery 1.1, bottom is Ng 1.1.*

Both speakers' F-patterns show typical values for a relatively back onset and a front offset, but the speakers differ considerably in the trajectories of all four formants, perhaps from the difference in time taken to say them. (So these two tokens differ a lot, but it must be remembered that it is the ratio of between- to within-speaker variance which is important). From the values of his F1, Ng, for example does not seem to change vowel height much. For this reason we have followed [1] in not trying to measure formant values at specific points, but in capturing the F-pattern trajectories with coefficients of formulae that describe the curves. We hope to get stronger LRs in this way. Two problems with the trajectory method, however, are that it may be influenced by different onset and offset perturbations caused by the consonants. For example, the typical bilabial transition in F2 caused by the following /p/ can be easily seen in figure 1. We do not have a problem at onset, as all the tokens occur after /s/. Here we took the onset of the F-pattern to be values at the first strong glottal pulse. But the tokens occur before different consonants and also word-finally in the corpus, and so this has to be controlled. We did this by taking as offset where F2 reaches its peak. The second problem also occurs with this sound: in telephone transmission, we expect low F1 values to be distorted, which means that the F1 values towards the end cannot be included which we cannot do if we model the F1

trajectory as a whole. It would be possible either to exclude F1 completely, or perhaps just include its intercept value, which models its value at onset.

The speech analysis software *Praat* was used to extract the formant centre-frequencies and F0 values. Wideband spectrograms were generated with a 0.005 sec. Gaussian window and a range from 0 to 5k, and formant tracks and F0 were superimposed on the spectrogram. The F1, F2 and F3 formant values as well as the F0 values were automatically tracked by *Praat*. For formants, the setting was determined by how well the tracking looked against the spectrogram. As a default, *Praat* was asked to track up to four formants with a formant ceiling of 4000 Hz, with the maximum setting having to be lowered to 3500 Hz for better tracking with some individuals. Formants were then estimated (Burg method, preemphasis from 50 Hz.), extracted, and input to Excel files manually for later analysis. F0, which was easier to extract, was similarly treated. The time setting was 0.01 sec. throughout. Poor extractions were either manually corrected or in really bad cases another token was chosen.

Code was written in *R* to model the formant and tonal F0 trajectories with polynomials (cubic for formants, quadratic for F0), and the coefficients were then input into the two-level kernel density multivariate likelihood ratio formula described in [6]. This is a formula from the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* which is used when it is suspected that the variables correlate. It is appropriate here because we would expect correlation between at least F2 and F3 in the high front rounded second diphthongal target. The likelihood ratio's numerator quantifies the similarity between the mean values of the offender and suspect, while the denominator quantifies the typicality of the difference against the reference population. The LR is the ratio of their values.

During the analysis for the current experiment, each trial for which a LR has to be estimated consists of the comparison of a set of values from one speaker in recording session one with the other set of values from the same speaker in recording session two; or with another set of values from another among the 14 other speakers. The difference between two speech samples under comparison is evaluated against a model of the distribution of same-speaker data to determine the probability of getting the difference assuming the two speech samples were spoken by the same speaker (the numerator of the LR); and against a model of the distribution of the other speakers to determine the probability of getting the difference assuming the two samples were spoken by different speakers (the LR denominator). In this experiment, because of the small number of speakers, we used leave-one-out cross-validation, but we have also included the results from an intrinsic comparison, where the speakers being compared are not removed from the reference sample.

F-pattern and F0 data from /ɔy/ were available from 15 speakers. This meant that 15 same-speaker comparisons, or target trials, and 105 different-speaker, or non-target trials, could be made. The raw MVLRs output from the test were then calibrated with logistic regression. It has been suggested that poor calibration is not likely to happen with traditional features from an analytically derived LR formula, as in the present study. But it may be the case that bad calibration comes simply from the number of variables used, which in this study is quite high: with coefficients from cubic polynomials from three formants we have 12 variables, and there are four variables for the F0. In order not to overload the MVLR formula, LRs were estimated for the F-pattern and F0 separately.

3. Results

3.1. Tippett plots

Results in forensic voice comparison experiments are usually presented with Tippett plots, which show the cumulative distribution of LRs from same-subject comparisons and different-subject comparisons, and show the proportion of LRs observed from same-speaker or different-speaker comparisons equal to or bigger than a given LR value. Based on this the trier of fact may see the probability of errors with the system used, which is also accordance with the *Daubert* criteria. In the kind of Tippett plot shown here, different-speaker (non-target) LRs decrease towards the right, and the same-speaker (target) LRs decrease towards the left.

The LR gives an estimate of strength of evidence: $\text{Log}_{10}\text{LR} < 0$ is hoped-for for a different-speaker hypothesis and $\text{Log}_{10}\text{LR} > 0$ for a same-speaker hypothesis. Given that, experimenters would hope the obtained strength estimate to correlate with the prosecution and defense hypothesis, rather than to give a strong counter-factual result. Therefore, they would want an evaluation of the system to penalize bad counter-factual LRs more than not-so-bad ones, and that information is what is provided by the log-likelihood ratio cost (C_{llr}). If the system is accurate, it should have C_{llr} values below 1, ideally a lot below 1. An additional measure of the system performance may be the equal error rate (EER). Strictly speaking the EER is not appropriate to likelihood ratio-based forensic voice comparison because it implies a posterior probability [5], but if we take the experiment as a test to see how well same-speaker speech samples can be distinguished from different-speaker speech samples it can be useful. Putting it simply, values for C_{llr} and EER are both the smaller, the better.

The top panel of figure 2 shows the Tippett plot for the calibrated non-cross-validated LRs for tonal F0 in /ɔy/. For non-cross-calibrated data, the C_{llr} is already 0.86, and the EER at ca. 40%. It will be probably worse for cross-calibrated comparison, which means that the tonal F0 is not delivering any information, or next to none, and indicates that the F0 in this Cantonese tone is not a useful feature in forensic voice comparison. On the other hand, the /ɔy/ F-pattern results in the bottom panel of figure 2 are a bit better, with a C_{llr} of 0.55 and an EER of about 20%, indicating that the /ɔy/ F-pattern would probably be of use in FVC. It is interesting to note the drastic narrowing effect of cross-validation in figure 2. The non-cross-validated data show a rather good performance with a substantially smaller C_{llr} value of 0.26. The main term governing the magnitude of the multivariate likelihood ratio is the ratio of between- to within-speaker variance, and it looks as if the side-to-side narrowing of the range from non-cross-validation (cross-validated and non-cross-validated LRs are very highly correlated) may be due to the reduction in between-speaker variance in the reference sample that occurs when one or two test speakers are removed from it.

3.2. Combination with Cantonese /iau/

The calibrated LRs from the tonal F0 are so bad that it was decided not to fuse them with the LRs from the F-pattern: at best they would give a minimal improvement and at worse a catastrophic fusion (i.e. a reduction in performance). To demonstrate the power of combining segments, however, we show the result of combining the LRs from the /ɔy/ F-pattern in this study with LRs from /iau/ in a separate study from the same 15 speakers [8]. The fusion was done with logistic-regression. The resulting Tippett is shown in figure 3. It can be

seen that there is an improvement on the individual sounds' performance, with a drop in C_{llr} to 0.44, and an EER of about 12%. The fused non-cross-validated comparisons are again very much better than the cross-validated.

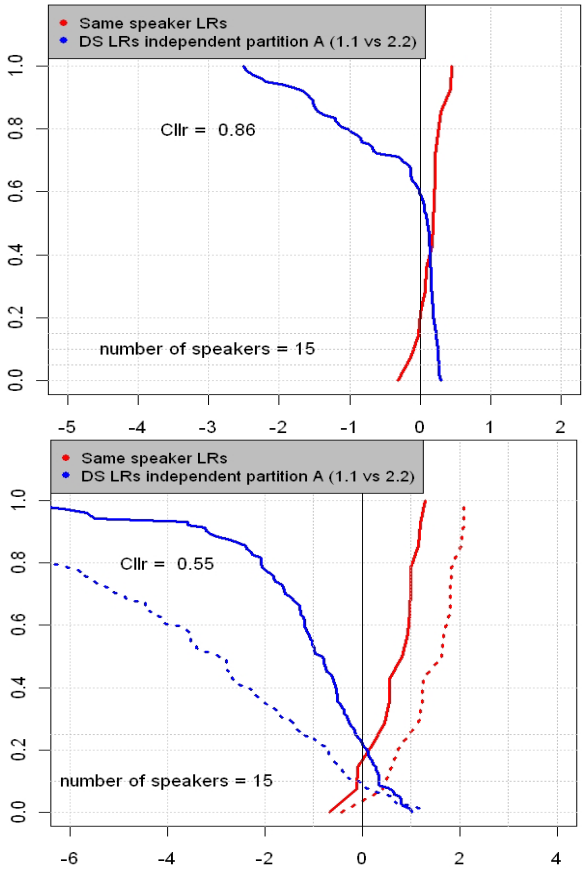

Figure 2: *Tippett Plots for calibrated LRs in* /ɔy/. *Top = non-cross-validated tonal F0 in Cantonese low-high rise tone. Bottom = F-pattern. x axis = log_{10}LR bigger than ... , y axis = cumulative proportion of different-speaker comparisons, 1- cum. prop. of same-speaker comparisons. Dotted line = non-cross-validated comparisons.*

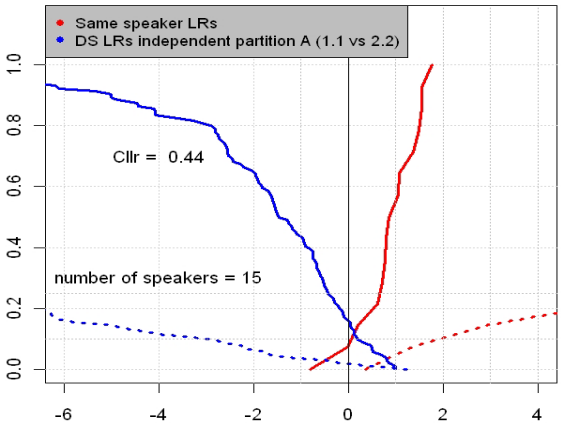

Figure 3: *Tippett Plot for Cantonese /ɔy/ & /iau/ non-contemporaneous LRs fused with logistic regression. Information as for figure 2.*

4. Summary & Conclusion

The aim of this paper was to see how well acoustic features of the Cantonese diphthong /ɔy/ perform in forensic voice comparison, using a kernel density multivariate likelihood ratio. F-pattern and tonal F0 trajectories were taken as the target features. 15 subjects were used with cross-validation with 7 to 10 tokens from each subject per recording. Bad results in terms of C_{llr} were obtained for the tonal F0 but not so bad for the F-pattern. An improvement in C_{llr} was demonstrated by fusing the LRs from /ɔy/ and /iau/.

Limitations are always hard to avoid and that applies to this study too. Since the current study was a pilot experiment, the number of speakers was restricted. Ideally an extrinsic comparison should be conducted with a much larger number of speakers so that a separate test and reference sample can be used and uncertainties concerning cross-validated results can be obviated. The recordings were clean, and well controlled. All of these factors will have the effect of increasing the C_{llr}. The upside of the study is that it shows the kind of improvement that can happen when features from more than one segment are used. It is possible, too, that we could also use features from the /s/ spectrum, which, because it is strongly labialized by the following /ɔ/, contains fairly low high amplitude frequency components that might be of use.

5. Acknowledgements

We would like to thank all the participants who have been recorded for the study, including the students in the UST and the CUHK, as well as our colleagues Boji, Jeffery, and Daniel who helped a lot in the experiment. 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*. The paper was written using findings from *Australian Research Council Discovery Grant* No. DP0774115. Many thanks too to our reviewers for their helpful comments and corrections.

6. References

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

[2] Rose, P. "Forensic Voice Comparison with Secular Shibboleths – a hybrid fused GMM-Multivariate likelihood-ratio-based approach using alveolo-palatal fricative cepstral spectra". Proc. International Conference on Acoustics Speech & Signal Processing, IEEE: 5900-5903, 2011.

[3] 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.

[4] 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.

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

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

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

[8] Chen, A., "Likelihood Ratio-based Forensic Voice Comparison with the Cantonese triphthong /iau/". Paper accepted for 14th Australasian International Speech Science & Technology Conference, Sydney, 2012.

[9] Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993) 113 S Ct 2786. 1993.