interspeech 2008 incorporating SST 08

Forensic Speaker Recognition in Chinese: A Multivariate Likelihood Ratio Discrimination on /i/ and /y/

Cuiling Zhang, [1,2] *Geoffrey Stewart Morrison,* [2] *Philip Rose* [2]

[1] Department of Forensic Science & Technology
China Criminal Police University, Shenyang, China
[2] School of Language Studies, Australian National University, Canberra, Australia
cuilingzhang@yahoo.com.cn, geoff.morrison@anu.edu.au, philip.rose@anu.edu.au

Abstract

A likelihood-ratio-based forensic speaker discrimination was conducted using the mean formant frequencies of Standard Chinese /i/ and /y/ tokens produced by 64 male speakers. The speech data were relatively forensically realistic in that they were relatively extemporaneous, were recorded over the telephone, and were from three non-contemporaneous recording sessions. A multivariate-kernel-density formula was used to calculate cross-validated likelihood ratios comparing all possible same-speaker and different-speaker combinations across sessions. Results were comparable with those previously obtained with laboratory speech in other languages. In general, greater strength of evidence was obtained for recording sessions separated by one week than for recording sessions separated by one month.

Index Terms: forensic speaker recognition, likelihood ratio, Chinese

1. Introduction

Forensic speaker recognition was first introduced in China in the 1980s. For more than twenty years forensic speech experts in China have employed so-called voiceprint identification, buttressed with some traditional auditory-phonetic approaches, to determine whether the offender voice and suspect voice come from the same speaker. The court usually only wants the expert testimony to be a bald "yes" or "no" to the question of the probability of the hypothesis, given the evidence, and regards conclusions containing any uncertainty as useless. This is true for fingerprinting, ballistics, handwriting, etc. as well as for speaker identification. It is well known that most offender and suspect samples cannot match perfectly because of within-speaker variation due to factors such as non-contemporaneity. Therefore, even if they are not one-hundred percent confident, experts usually have to take certain risks in order to present a definitive conclusion which satisfies the present standards for admissibility in court.

Over approximately the last decade, some researchers have been exploring a more logically and legally correct way of evaluating the strength of forensic speaker recognition evidence in the form of the likelihood ratio of Bayes' theorem [1,2]. It has been shown that this approach can emulate the methodology for the evaluation of DNA evidence [3]. The strength of evidence is estimated with a likelihood ratio. This is the ratio of the probabilities of the evidence under the competing hypotheses, i.e., the probability of obtaining the observed differences between the offender and suspect speech samples assuming they were produced by the same speaker relative to the probability of obtaining the observed differences assuming they were produced by different speakers. Likelihood ratio values greater than unity (or log likelihood ratios greater then zero) support the prosecution hypothesis that the suspect is the source of the incriminating speech, and likelihood ratio values less than unity (or log likelihood ratios less then zero) support the defense. The magnitude of the likelihood ratio is proportional to the strength of the evidence, with values close to unity meaning that the evidence is just about as likely under both hypotheses, and therefore useless. Only likelihood ratio values much larger or much smaller than unity (log likelihood ratios of large magnitude) are of judicial value.

To our knowledge, this paper is the first to apply a likelihood-ratio approach to forensic speaker recognition in Chinese. Although the effectiveness of the approach used here, that is, with traditional as opposed to automatic features, has been tested in several studies on Australian English and Japanese, e.g. [4,5,6], none of these studies have made use of speech data that is truly realistic. The data analyzed in the present paper is more forensically realistic in that it was collected via multiple non-contemporaneous telephone interviews with participants speaking in a relatively extemporaneous manner. This research is part of a larger investigation, funded by the *International Centre for Excellence in Asian-Pacific Studies*, into the effectiveness of the likelihood ratio approach applied to Chinese. The paper constitutes an initial assessment of the forensic discriminability of the mean formant frequencies of two vowel phonemes /i/ and /y/, using cross-validated multivariate likelihood ratios obtained via a formula developed by Aitken and Lucy at the *Joseph Bell Centre for Forensic Statistics and Legal Reasoning* [7].

2. Methodology

2.1. Data collection

Speech data were collected from 64 young male speakers of Standard Chinese. All speakers had grown up in the city of Shenyang, they were aged between 19 and 23, and were students at the *China Criminal Police University*. They were recorded three times, with the second and third recording sessions being approximately one week and one month after the first session. Speech was elicited by research assistants, who were a couple of years older than the speakers. The research assistants phoned the subjects and asked them a number of questions such as: "What's your name?", "What's your mobile number?", "What's the address of your university?"

Recordings were made via the university internal telephone system using a *KCM* HCD9999P/TSDL telephone, which has an in-build analogue cassette tape recording facility. Recordings were digitised using a *SANYO* DC-PT70 DVD Micro Component System, and saved as 16 bit PCM sound files at a sampling frequency of 11.025 kHz.

2.2. Acoustic measurement

For each session of each speaker, ten stressed tokens of /i/ and six stressed tokens of /y/ were identified and measured. The operational criterion for 'stressed' was that the vowel have a duration greater than 40 ms. The first three formants (F1, F2, and F3) were measured using *Praat* [8]. Formant frequencies were calculated using the Burg LPC method looking for four formants below 4 kHz. The mean frequencies over the timecourse of the vowel were recorded.

2.3. Statistical analysis

In order to calculate likelihood ratios, F1, F2, and F3 values were entered into the multivariate kernel density formula described in [7] and implemented in [9]. This formula assesses the difference between suspect and offender samples with respect to their typicality in reference to a background distribution estimated using data from a sample taken from the appropriate population. Within-speaker variation is estimated assuming a normal distribution, and between-speaker variation is estimated with a Gaussian kernel density model.

For both /i/ and /y/, cross-validated likelihood ratios were then calculated for all same-speaker and different-speaker pairs in the data: Within each speaker's data set, session 1 was compared with session 2, session 1 with session 3, and session 2 with session 3. Each speaker's session 1 was compared with every other speaker's sessions 1, 2, and 3; each speaker's session 2 was compared with every other speaker's sessions 1, 2, and 3; and each speaker's session 3 was compared with every other speaker's sessions 1, 2, and 3. For each comparison cross-validation was used, whereby the background distribution was calculated using the data from all the speakers except for the speakers being compared (in calculating the background distribution, data from all three sessions were pooled within speakers).

The bandpass properties of telephone systems typically make some F1 measurements unreliable, in particular those of high vowels such as /i/ and /y/. Therefore, in addition to conducting analyses using all three formants, a second set of analyses was conducted using only F2 and F3 measurements. This provides additional realism.

3. Results

The cross-validated likelihood ratios for /i/ and /y/ are presented in Figures 1 and 2 respectively in the form of Tippett plots [10,11]. The Tippett plots show the cumulative proportions of cross-validated same-speaker and different-speaker log likelihood ratios. Curves rising to the left represent the proportion of different-speaker comparisons with $\log_{10}$ likelihood ratios equal to or greater than the value indicated on the *x*-axis. Curves rising to the right represent the proportion of same-speaker comparisons with $\log_{10}$ likelihood ratios equal to or less than the value indicated on the *x*-axis. An ideal likelihood-ratio based forensic speaker recognition system would return log likelihood ratios which are much larger than zero if the two speech samples are produced by the same speaker, and log likelihood ratios much smaller than zero if the two speech samples are produced by different speakers. For the cross-validated likelihood ratios reported here, we know whether each comparison was from two sessions from the same speaker or a comparison of data from different speakers. The sign and magnitudes of these cross-validated log likelihood ratios therefore provide a measure of the effectiveness of the system (i.e., the formula and the features).

Figure 1 represents the results of the /i/ analyses. The three-formant and two-formant analyses had equal-error rates of 27.6% and 26.1% respectively. It can be seen that excluding F1 measurements resulted in somewhat less negative different-speaker log likelihood ratios, and therefore better results for different speaker comparisons. In contrast, excluding F1 did not result in consistently more positive same-speaker log likelihood ratios. The strength of evidence for same-speaker pairs was, in fact, somewhat limited: The largest log likelihood ratio obtained in the three-formant analysis was 2.397, indicating that one would be 249 times more likely to observe the mean formant differences between the speech samples under the hypothesis that they were produced by the same speaker than under the hypothesis that they were produced by different speakers. In terms of the verbal scale for likelihood ratios proposed in [12], this would be considered "moderately strong" evidence in support of the same-speaker hypothesis. The remainder of the same-speaker comparisons produced lower log likelihood ratios. The worst result for a same-speaker comparison in the three-formant analysis was a log likelihood ratio of –5.698, indicating that one would be 498 790 times more likely to observe the formant differences between the speech samples under the hypothesis that they were produced by the different speakers than under they hypothesis that they were produced by the same speakers. Contrary to fact this provides "very strong evidence" [12] in support of the different speaker-hypothesis.

Figure 2 shows the results of the /y/ analyses. As can be seen, results were generally similar to those obtained for /i/. The three-formant and two-formant analyses had equal-error rates of 25.1% and 27.7% respectively.

Figure 3 shows Tippett plots of the results of a Naïve Bayes combination of /i/ and /y/ likelihood ratios made under the assumption that the two sets of data are uncorrelated (an assumption that is also phonetically naïve!). It can be seen that a substantial improvement in log likelihood ratio values was obtained, with an equal-error rate of between ca. 20% and 25%. As implied by "naïve", the assumption of zero correlation between /i/ and /y/ formant patterns does not hold, and the results of the Naïve Bayes combination, although representing a considerable discriminatory improvement, are suspect both statistically and forensically. It is interesting to note that correlation between the actual likelihood ratios, however, was relatively low: 0.124 for the two-formant model, but only 0.037 for the three-formant model.

Figures 4 and 5 provide Tippet plots which show the effect of elapsed time on the likelihood ratios derived from /i/ and /y/ respectively. They compare the results of analyses conducted using only data from recording sessions 1 and 2, which were separated by one week (a relatively short elapsed time), and the results of analyses conducted using only data from recording sessions 1 and 3, which were separated by one month (a relatively long elapsed time). As expected, greater discrimination power was obtained using samples from recordings separated by one week than using samples from recordings separated by one month. This is probably due to the well-known effect of within-speaker variability increasing with the elapsed time between sessions.

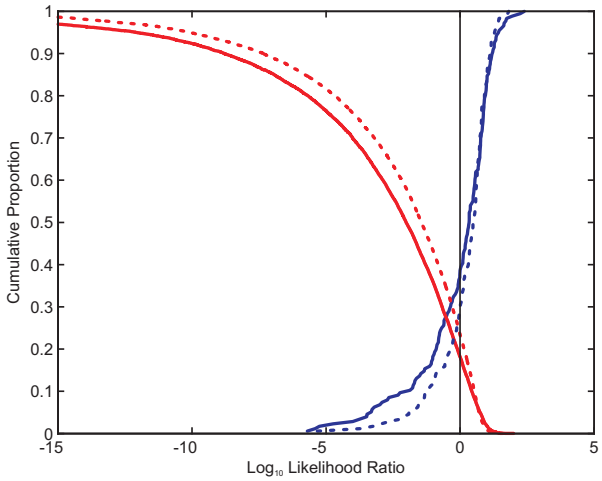

Figure 1 *Tippett plot of cross-validated multivariate likelihood ratios for /i/. Solid lines: using F1, F2, and F3. Dotted lines: using F2 and F3 only.*

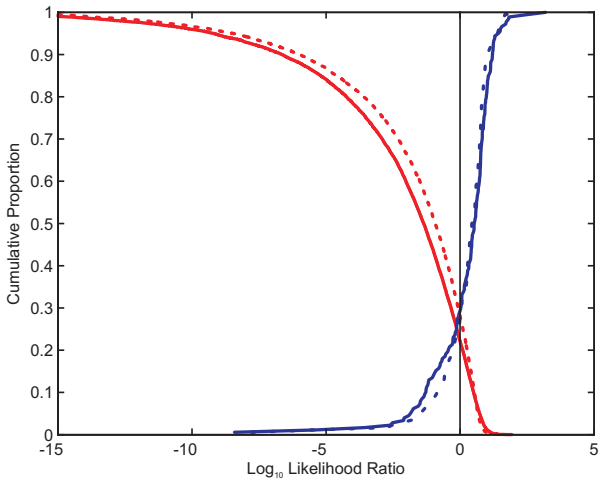

Figure 2 *Tippett plot of cross-validated multivariate likelihood ratios for /y/. Solid lines: using F1, F2, and F3. Dotted lines: using F2 and F3 only.*

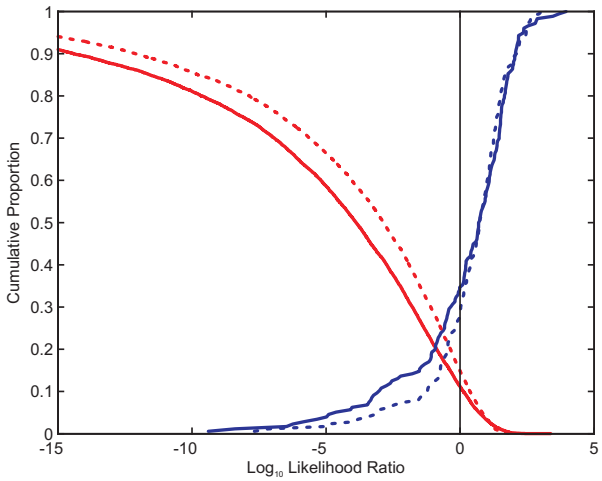

Figure 3 *Tippett plot of cross-validated multivariate likelihood ratios for Naïve Bayes combination of /i/ and /y/. Solid lines: using F1, F2, and F3. Dotted lines: using F2 and F3 only.*

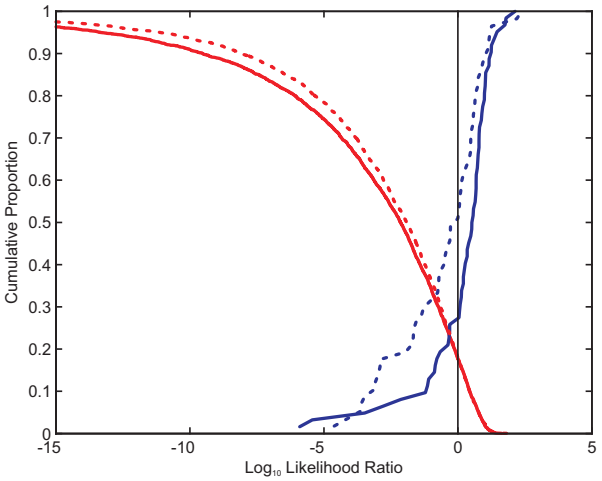

Figure 4 *Tippett plot of cross-validated multivariate likelihood ratios for /i/ using F1, F2, and F3 Solid lines: one week between recording sessions. Dotted lines: one month between recording sessions.*

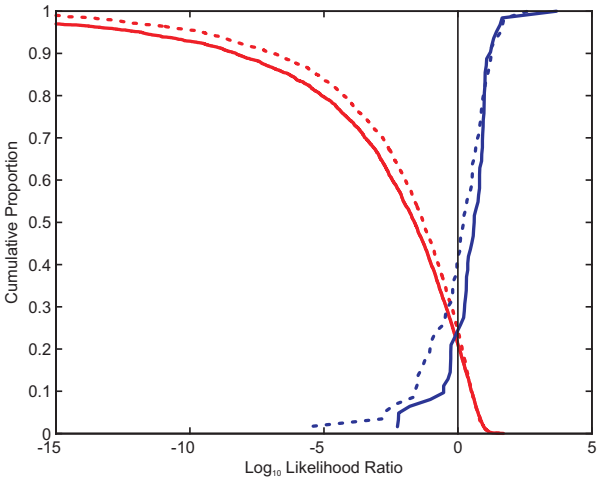

Figure 5 *Tippett plot of cross-validated multivariate likelihood ratios for /y/ using F1, F2, and F3 Solid lines: one week between recording sessions. Dotted lines: one month between recording sessions.*

4. Discussion and Conclusion

This paper has presented results of a likelihood-ratio-based forensic speaker discrimination using the formant patterns of two Standard Chinese vowels. Although the cross-validated likelihood ratios were not particularly impressive, they were similar to results obtained in [6] which used very similar procedures to obtain likelihood ratios based on Australian English vowels. It is also worth noting that whereas the speech data in [6] was read speech recorded in a laboratory setting, the present study made use of speech collected using a more forensically realistic procedure, i.e., relatively extemporaneous, non-contemporaneous telephone speech from 64 speakers. We may conclude therefore that a certain amount of forensic discriminability can be expected from vocalic formant patterns, even under forensically realistic conditions.

There are several ways in which it may be possible to improve on the results reported here:

Only ten tokens of /i/ and six of /y/ were measured for each speaker in each session. The recordings include additional tokens of these vowels, and it should be possible to

measure approximately twice the number of tokens used here. A larger amount of data will allow for better estimates of the within- and between-speaker probability density functions, and therefore better estimates for likelihood ratios.

Data from additional vowels such as /a/, /ə/, and /u/ are also available in the recordings, and combining data from a larger number of vowel categories may also lead to further improvement in discrimination.

Stressed Chinese vowels are pronounced with a voice source which encodes tone, and different tones may affect the vocalic formant pattern. The analyses reported here did not take account of tone, and it may be possible to achieve better results if tone were controlled for. Tonal fundamental frequency itself may also provide additional evidence relevant for speaker differentiation.

In the present study mean formant values were used as features. In an as-yet unpublished study from our lab [13] which fitted parametric curves to formant trajectories, we found a substantial improvement in performance, compared to a model which used formant values from putative initial and final targets of the vowel. It may be that additional discriminatory power can also be extracted from Chinese vowels using a formant-trajectory approach.

To our knowledge, this is the first report of a likelihood-ratio-based forensic speaker discrimination conducted on Chinese. We also believe that it is the first report of a likelihood-ratio-based forensic speaker discrimination using traditional features where the speech data were forensically realistic and collected from a moderately large number of speakers. It is hoped that additional research on likelihood-ratio approaches to forensic speaker recognition, and other forensic technologies, will lead both to improved results, and the adoption of likelihood-ratio approaches in forensic casework in China, with positive results for the justice system.

5. Acknowledgements

This research was supported by a grant from the *International Centre for Excellence in Asian-Pacific Studies* at the *Australian National University*. We thank all our informants and research assistants for their invaluable assistance.

6. References

[1] Aitken, C. G. G, and Taroni, F., Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, Chichester, UK, 2004.

[2] Rose, P., Expert Evidence, Issue 99: The Technical Comparison of Forensic Voice Samples. Thomson, Sydney, Australia, 2003.

[3] González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D. T., and Ortega-García, J. "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition", IEEE Trans. on Audio, Speech, and Lang. Proc., Vol. 15, 2007, 2104–2115.

[4] Alderman, T. G., Forensic Speaker Identification: A Likelihood Ratio-Based approach Using Vowel Formants. Lincom, Munich, Germany, 2005.

[5] Rose, P., Osanai, T., & Kinoshita, Y., "Strength of forensic speaker identification evidence: Multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold", Forensic Linguistics, Vol. 10, 2003, 1350–1771.

[6] Rose, P., "Forensic speaker discrimination with Australian English vowel acoustics", Proceedings of the 16th International Congress of Phonetic Sciences, Universität des Saarlandes, Saarbrücken, Germany, 2007, 1817–1820.

[7] Aitken, C. G. G, and Lucy, D., "Evaluation of trace evidence in the form of multivariate data", Applied Statistics, Vol. 54, 2004, 109–122.

[8] Boersma, P, & Weenick, D., Praat: Doing phonetics by computer [software], 2008. Available: http://www.praat.org

[9] Morrison, G. S., Matlab Implementation of Aitken & Lucy's (2004) Forensic Likelihood-Ratio Software Using Multivariate-Kernel-Density Estimation [software], 2007. Available: http://geoff-morrison.net

[10] González-Rodríguez, J., Drgajlo, A., Ramos-Castro, D., García-Gomar, and Ortega-García, J., "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition", Comp. Lang. and Speech, Vol. 20, 2006, 331–335.

[11] Rose, P., "Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination", Proceedings of the Odyssey Speaker and Language Recognition Workshop, 2006, 1–8.

[12] Champod, C., and Evett, I. W., "Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification'", Forensic Linguistics, Vol. 7, 2000, 1350–1771.

[13] Morrison, G. S., "Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/", Manuscript submitted for publication.