Forensic comparison of Chinese female voices

Cuiling Zhang
Department of Criminal Science & Technology, China Criminal Police University

Geoffrey Stewart Morrison
School of Language Studies, Australian National University

http://forensic-voice-comparison.net

Research Project Funded by
International Association for Forensic Phonetics and Acoustics Research Grant

http://www.iafpa.net/research.htm

Brief summary of research project:
This project will investigate the effectiveness of forensic voice comparison based on the formant trajectories of Chinese vowels spoken by adult female speakers of Standard Chinese. It will also investigate the effect of varying the size of the population sample. Results will be assessed using the log-likelihood-ratio cost function (C_{llr}).

Anticipated Duration of Project:
13 months

Breakdown of Research Expenses:

Salaries and honoraria:	
participant payments: 40 Yuan × 2 sessions × 60 participants =	4800 Yuan
research assistant: 50 Yuan per hour × 90 hours =	4500 Yuan
Recording equipment:	
2 × Sennheiser MKE 2 P-C Microphones	6400 Yuan
1 × Edirol UA-24EX Soundcard	2000 Yuan
2 × microphone clips and cables	700 Yuan
Total :	18400 Yuan ≈ 2000 Euro

Statement of proposed research:

1. Background

1.1 Motivation

Research on forensic voice comparison has primarily focused on male voices – within the last five years none of the papers published in the *International Journal of Speech, Language and the Law* have focused on female voices. This is understandable given that most casework deals with male voices; however, in the Principal Applicant's casework experience approximately 20% of cases involve female voices. Some forensic voice comparison research on Chinese has been published (Zhang & Rose, 2007; Zhang, Morrison, & Rose, 2008; Zhang, & Rose, 2008), but only on male voices. There is therefore a need to conduct research on the effectiveness of forensic techniques when applied to female voices.

A promising line of investigation for forensic voice comparison extracts information from parametric curves fitted to formant trajectories (McDougall, 2006; McDougall & Nolan, 2007; Morrison, 2008a, in press, Morrison & Kinoshita, 2008). Simple two-point models have been found to be effective for vowel phoneme identification (Adank, van Hout, and Smits 2004; Andruski & Nearey 1992; Hillenbrand, Clark, & Nearey 2001; Hillenbrand, Getty, Clark, & Wheeler 1995; Hillenbrand and Nearey 1999). There are therefore potentially numerous degrees of freedom in the exact trajectory between the points relevant for phoneme identification which may convey information relevant to speaker identity. If only initial and final formant values are relevant for phoneme identity, then on the basis of physiological and motor learning idiosyncrasies different speakers could potentially produce consistently different trajectory between the two points. The present study will therefore examine the effectiveness of forensic voice comparison based on parametric curves fitted to the formant trajectories of some frequently-occurring Chinese diphthongs and triphthongs.

1.2. Objectives

– To collect a database of recordings of Standard-Chinese female voices which can be used for forensic voice comparison research in the present and subsequent projects (the database will be made available to other researchers), and which can be used in casework.

– To determine which of the following combinations of vowel phonemes, parametric representations of formant trajectories, and likelihood-ratio calculation procedure are the most effective for forensic comparison of Standard-Chinese female voices.

> - vowel phonemes / iao1, uo3, ao4, iou4 / (these phonemes were selected because, on the basis of an earlier study on male voices, tokens of these phonemes were expected to exhibit substantial formant movement and to have relatively high frequency of occurrence)
>
> - polynomials, discrete cosine transformations
>
> - multivariate-kernel-density model (Aitken & Lucy, 2004), Gaussian mixture models (Reynolds & Rose, 1995)

– To determine the number of speakers necessary in the population sample in order to achieve asymptotic behavior in the likelihood ratios.

2. Methodology

2.1. Speech-Sample Collection

We plan to collect speech samples from 60 adult female speakers of Standard Chinese. Each speaker will be recorded on two occasions separated by approximately two weeks (this will allow for between-speaker as well as within-speaker comparisons). The recording procedure described below is based on a procedure adopted for the collection of the Spanish Guardia Civil’s Baeza database.

Two speakers who know each other will be recorded holding a telephone conversation. In order to elicit speech, the participants will be given several tasks to perform, some designed to elicit particular speech items and others freer. For example, in order to elicit numbers one participant will be given a poorly reproduced fax transmission including telephone numbers and prices, and to check the numbers they will have to phone the other participant who has the original document. Conversations will last around 15 minutes in order to obtain at least 5 minutes of speech per participant.

Each speaker will sit in a different room and hold a telephone conversation via the internal telephone system. Each speaker will be wear a high-quality flat-frequency-response lapel microphone connected to a soundcard and computer in a third room. The signal from one microphone will be recorded on one recording channel and the signal from the other microphone on another channel.

If one begins with high quality recordings, these can potentially be filtered or directly passed through different transmission channels to examine the effect of transmission channel on system performance. This is not possible if one begins with poor quality recordings. For the present study we will be focusing of second and third formant trajectories which are usually assumed to be relatively robust to transmission channel effects (an assumption that is potentially testable if one begins with high quality recordings). That said, formant tracking is generally harder and more care must be taken with poorer quality recordings. If one assumes that it is true that the formants are robust to transmission channel effects, it is therefore more efficient to track the formants of high quality recordings.

2.3. Acoustic Measurement

Stressed tokens of / iao1, uo3, ao4, iou4 / in the speech samples will be manually marked and labeled. These are relatively frequently occurring phonemes and we expect to be able to extract around ten tokens of each phoneme from 5 minutes of speech. The trajectories of the frequencies of the first three formants will be tracked. These tasks will be achieved using software previously developed by the Second Applicant (Morrison, 2008b, Morrison & Nearey, 2008). The software is designed to be ergonomically efficient allowing the user to rapidly mark, label, and measure a large number of vowels. The formant tracking software is semi-automatic, and the task of the user is to check the results of the automatic procedure. Formant tracks measured using different linear-predictive-coding parameters are available, and the software automatically picks the best tracks on the basis of a series of heuristics. The software displays alternative formant tracks overlaid on the spectrogram, and also synthesizes a vowel on the basis of the selected formant tracks. If user intervention is necessary, it is often a simple matter of clicking on one of the alternative format tracks. If hand editing is necessary, this can also be achieved using a point-and-click procedure. In previous research projects we have been able to mark and label around 300 tokens per day, and measure around 1000 tokens per day.

2.4. Information Extraction

Information extraction, and calculation and evaluation of likelihood ratios will replicate and extend procedures adopted in Morrison (in press) and Morrison & Kinoshita (2008). Parametric curves (polynomials and discrete cosine transforms) will be fitted to the trajectories of the second and third formant (the first formant will be excluded since it would normally be expected to be compromised by the bandpass properties of a regular telephone transmission system). The coefficient values from the parametric curves will be used as variables in the calculation of likelihood ratios. Different types and orders of parametric curves, and different scalings of time and frequency will be tested in order to determine which is most effective for forensic speaker comparison.

2.5. Calculation of Likelihood Ratios

Several methods for the calculation of likelihood ratios will be compared, including the multivariate-kernel-density model of Aitken & Lucy (2004), and Gaussian mixture models (Reynolds & Rose, 1995). Cross-validated likelihood ratios will be calculated for each possible same-speaker and different-speaker pair. Separate sets of likelihood ratios will be calculated for each vowel phoneme.

2.6. Evaluation, Calibration, and Fusion

The cross-validated likelihood ratios obtained using different systems (combinations of parametric curves and methods for the calculation of likelihood ratios) will be evaluated using the log-likelihood-ratio cost (C_{llr}) (Brümmer & du Preez, 2006). Each system will be calibrated, and likelihood ratios sets derived from different vowel phonemes will be fused using logistic regression (Brümmer *et al*., 2007; Pigeon, Druyts, & Verlinde, 2000). C_{llr} will be used to determine which individual-phoneme systems and which fused-systems are most effective for forensic voice comparison.

2.7. Population Sample Size

A series of Monte-Carlo-type simulations will be conducted in order to investigate the performance of the forensic-voice-comparison systems as the number of speakers in the population sample is varied (a similar investigation, but using randomization-type simulations rather than Monte-Carlo-type simulations was conducted by Ishihara & Kinoshita, 2008). A population model will be estimated using Gaussian mixture models trained on all the data extracted from the voice database. This model will then be treated as if it is a true model of a simulated population and used to generate samples of *m* tokens from *n* speakers. Although the model is trained on a finite amount for data from a finite number of speakers, it can be used to generate an infinite amount of simulated data from an infinite number of simulated speakers. Multiple sets of simulated samples of speaker size *n* will be generated and used as the reference sample for likelihood ratio calculation. Systematic probe sets of known and questioned voice samples will be used, the sets covering a range of different similarities and typicalities in the sample space. Of particular interest is the question of the number of speakers *n* necessary to obtain asymptotic behavior – a practical indicator of the number of speakers needed for future research and casework. Since the simulated population is represented by a parametric model, it is also possible to calculate the theoretically correct likelihood ratio for each pair of probes and therefore also measure the accuracy as well as the precision of the likelihood ratios calculated by the systems using different *n*.

3. Research Rôles

Research Assistant:
  Participant recruitment, scheduling, and data collection (90 hours)
Principal Applicant:
  Supervision of Research Assistant, marking and labeling, paper writing
Second Applicant:
  Formant measurement, statistical analysis, paper writing

4. Timeline

2009-03 – 2009-05 Data collection
2009-06 – 2009-11 Acoustic analysis
2009-12 – 2010-01 Statistical analysis
2010-02 – 2010-03 Paper writing

5. Measurable Outcomes

On the basis of this project we plan to write a research paper and submit it for publication in a referred journal such as *Forensic Science International*, or the *International Journal of Speech, Language and the Law*. We also plan to present the results at the 2010 meeting of a major international conference such as *Interspeech*, *Acoustical Society of America*, or *International Association for Forensic Phonetics and Acoustics*.

6. Additional Research

Following the proposed research project described above, we plan to use the Chinese female voice database to conduct additional research. We plan to replicate the methodology of the proposed research project applied to other vowel phonemes. We also plan to investigate the effectiveness of fitting parametric models to the fundamental frequency trajectories of different tones on different Chinese vowels. As soon as the database has been collected, we will share it with collaborators at the Autonomous University of Madrid and the University of New South Wales, who will conduct automatic forensic voice comparisons. We will potentially also make use of the database to investigate the effectiveness of procedures such as automatic formant tracking, and forced alignment of automatic speech recognition output with orthographic transcription to locate phonemes and phoneme boundaries.

References

Adank, P. van Hout, R., and Smits R. (**2004**) "An acoustic description of the vowels of Northern and Southern Standard Dutch," J. Acoust. Soc. Am. **116**, 1729–1738.

Aitken, C. G. G., and Lucy, D. (**2004**). "Evaluation of Trace Evidence in the Form of Multivariate Data," App. Stat. **54**, 109–122.

Andruski, J. E. and Nearey, T. M. (**1992**) "On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables," J. Acoust. Soc. Am. **91**, 390–410.

Brümmer, N., Burget, L., Cernocký, J. H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., and Strasheim, A. (**2007**). "Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006," IEEE Trans. on Audio, Speech and Lang. Proc. **15**, 2072–2084.

Brümmer, N. and du Preez, J. (**2006**). "Application Independent Evaluation of Speaker Detection," Comput. Speech Lang. **20**, 230–275.

Hillenbrand, J. M., Clark, M. J., and Nearey, T. N. (**2001**) "Effect of consonant environment on vowel formant patterns," J. Acoust. Soc. Am. **109**, 748–763.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**) "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hillenbrand, J. M. and Nearey, T. N. (**1999**) "Identification of resynthesized /hVd/ syllables: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.

Ishihara, S., & Kinoshita, Y. (2008) "How many do we need? Exploratyion of the population size effect on the performance of forensic speaker classification," *Proceedings of Interspeech 2008 Incorporating SST 2008, Brisbane, Australia* (International Speech Communication Association), 1941–1944.

McDougall, K. (**2006**) "Dynamic features of speech and the characterisation of speakers," Int. J. Speech, Lang. Law **13**(1), 89–126.

McDougall, K. and Nolan F. (**2007**) "Discrimination of speakers using the formant dynamics of /u/ in British English," *Proceeding of the 16th International Congress on Phonetic Sciences, Saarbrücken*, 1825–1828.

Morrison, G. S. (**2008a**). “Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/,” Int. J. Speech, Lang. Law **15**(2).

Morrison, G. S. (**2008b**). SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files (Release 2008-12-19). [Computer software. Available: http://geoff-morrison.net]

Morrison, G. S. (**In Press**). “Likelihood-ratio forensic speaker comparison using parametric representations of the formant trajectories of diphthongs,” J. Acoust. Soc. Amer. **124**(4).

Morrison, G. S., and Kinoshita, Y. (**2008**) “Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories,” *Proceedings of Interspeech 2008 Incorporating SST 2008, Brisbane, Australia* (International Speech Communication Association), 1501–1504.

Morrison, G. S., and Nearey, T. M. (**2008**). FormantMeasurer: Software for efficient human-supervised measurement of format trajectories (Release 2008-12-21). [Computer software. Available: http://geoff-morrison.net]

Pigeon, S., Druyts, P., and Verlinde. P. (**2000**). “Applying logistic regression to the fusion of the NIST’99 1-speaker submissions,” Digit. Sig. Proc. **10**, 237–248.

Reynolds, D. A., and Rose, R. C. (**1995**). “Robust test-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Process. **3**, 72–83.

Zhang, C., Morrison, G. S., and Rose, P. (**2008**). “Forensic speaker recognition in Chinese: A multivariate likelihood ratio discrimination on /i/ and /y/,” *Proceedings of Interspeech 2008 Incorporating SST 2008, Brisbane, Australia* (International Speech Communication Association), 1937–1940.

Zhang, C., and Rose P. (**2007**). “Strength evaluation of forensic speaker recognition evidence in Chinese with a Bayesian Likelihood Ratio”, Paper presented at the Australian Research Council Human Communications Network Workshop on Forensic Speaker Recognition (FSI not CSI: Perspectives in State-of-the-Art Forensic Speaker Recognition), Sydney, New South Wales, Australia.

Zhang, C., and Rose P. (**2008**) “ 关于法庭语音证据力度的评价 Evaluation of forensic speaker recognition evidence],” Paper presented at the 8th Phonetic Conference of China and the International Symposium on Phonetic Frontiers, Beijing, China.