FORENSIC VOICE COMPARISON USING CHINESE /iau/

Cuiling Zhang[a,b], *Geoffrey Stewart Morrison*[b] *& Tharmarajah Thiruvaran*[b]

[a]Department of Forensic Science & Technology, China Criminal Police University, Shenyang, China;
[b]Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia
cuilingzhang@yahoo.com.cn; geoff-morrison@forensic-voice-comparison.net; thiru@ee.unsw.edu.au

ABSTRACT

An acoustic-phonetic forensic-voice-comparison system extracted information from the formant trajectories of tokens of Standard Chinese /iau/. When this information was added to a generic automatic forensic-voice-comparison system, which did not itself exploit acoustic-phonetic information, there was a substantial improvement in system validity but a decline in system reliability.

Keywords: forensic voice comparison, acoustic-phonetic, automatic

1. INTRODUCTION

A number of studies, e.g. [5, 10, 14, 19] conducted within the new paradigm for forensic-comparison science [9, 12] have explored the effectiveness of acoustic-phonetic forensic voice comparison based on the coefficient values of parametric curves fitted to the formant trajectories of diphthongs. The earlier studies used controlled (e.g., read) speech, and did not investigate whether this technique lead to improvements over an automatic system. It is important to ascertain whether the expense of the human labor involved in acoustic-phonetic procedures is justified by substantial improvement in performance over a cheaper automatic system.

The present paper explores the effectiveness of the formant trajectory technique applied to tokens of the Standard Chinese triphthong /iau/ on tone 1 (high level), occurring in the word 一 *yao* "one". Triphthongs being more complex than diphthongs, it may be possible to extract more useful information from a triphthong than a diphthong. The tokens are extracted from a database of spontaneous speech, being more forensically realistic in this respect than the controlled speech of earlier studies. The effectiveness of the formant-trajectory technique is assessed as the improvement in performance obtained when the latter is added to a generic automatic forensic-voice-comparison system. Improvement in performance is measured using the log-likelihood-ratio cost (C_{llr}) as a metric of validity [4, 7] and a parametric estimate of the 95% credible interval (CI) for the likelihood ratios (LRs) as a metric of reliability [15, 19, 20].

2. METHODOLOGY

2.1. Data

The data were extracted from a database [23] of voice recordings of female speakers of Standard Chinese (a.k.a. Mandarin and Putonghua). See [17] for details of the data collection protocol. The data consisted of 2 recordings of each of 60 speakers. The speakers were all first-language speakers of Standard Chinese from northeastern China, and were aged from 23 to 45 (with most being between 24 and 26). The recordings used were from an information exchange task conducted over the telephone: Each of a pair of speakers received a "badly transmitted fax" including some illegible information, and had to ask the other speaker to provide them with the missing information. The original recordings were approximately 10 minutes long, with the second recording of each speaker recorded 2-3 weeks after the first. Recordings were high quality, recorded at 44.1 kHz 16 bit using flat-frequency response lapel microphones. The present paper should be considered a preliminary to testing more forensically realistic conditions including transmission-channel and speaking-style mismatches.

Data from the first 20 speakers (01-04, 09-20, 22, 25, 26, 28) were used as background data, data from the next 20 speakers (29-48) were used as development data, and data from the last 20 speakers (49-68) were used as test data.

2.2. Acoustic-phonetic system

There were between 8 and 20 stressed tokens of /iau/ per speaker per recording. The tokens were manually located and marked using *SoundLabeller* [13], and the trajectories of the first three formants (F1, F2, F3) of each token were measured using *FormantMeasurer* [16]. Discrete cosine transforms (DCTs) were fitted to each formant trajectory and the DCT coefficient values were used to calculate likelihood ratios.

After tests on the development set using different numbers of DCT coefficients, and different combinations of formants, the zeroth through fourth DCT coefficients fitted to the F2 and F3 trajectories were used in the final analysis.

Likelihood ratios were calculated using the multivariate kernel density (MVKD) formula [1, 8]. The background data were used to model the distribution of the features in the population. Likelihood ratios were calculated for each pair of speakers in the development data: Each speaker's first recording (nominal offender recording) was compared with their own second recording and with every other speaker's first and second recording (nominal suspect recordings), resulting in 20 likelihood ratios from same-speaker comparisons and 760 pairs of likelihood ratios from different-speaker comparisons. The likelihood ratios from the development set were used to calculate weights for logistic-regression calibration [2, 4, 7, 11]. The pooled procedure for the calculation of the calibration weights [18] was adopted (for both the acoustic-phonetic and the automatic system this resulted in both greater precision and greater accuracy). The MVKD procedure was then applied to the test data to obtain 20 likelihood ratios from known same-speaker comparisons and 760 pairs of likelihood ratios from known different-speaker comparisons, and these were calibrated using the weights which had been calculated using the likelihood ratios from the development set.

2.3. Automatic system

The automatic forensic-voice-comparison system was of generic design. 16 mel-frequency-cepstral-coefficient (MFCC) values were extracted every 10 ms over the entire speech-active portion of each recording using a 20 ms wide hamming window. Delta coefficient values were also calculated and included in the subsequent statistical modeling [6]. A Gaussian mixture model - universal background model (GMM-UBM) [22] was built using the background data to train the background model. After tests on the development set using different numbers of Gaussians, the number of Gaussians used for testing was 1024. For each comparison pair, the nominal suspect recording was used to build a suspect model and the nominal offender recording was used as probe data to calculate a score. As with the acoustic-phonetic system, scores were calculated for comparison pairs in the development data and these were used to calculate calibration weights which were used to calibrate the scores from the test set and convert them into calibrated likelihood ratios.

2.4. Fused system

The test scores from the acoustic-phonetic and automatic system were fused using logistic-regression fusion [2, 3, 4, 11, 21, 22]. As was the case for the calibration of the individual systems, fusion weights were calculated using scores from the development set and then applied to the scores from the test set.

3. RESULTS

The validity and reliability measures on the test results from the acoustic-phonetic, automatic, and fused systems are given in Table 1. The 95% credible intervals were calculated using the parametric procedure on different-speaker pairs, e.g., Speaker 01 Recording A versus Speaker 02 recording B, and Speaker 02 Recording A versus Speaker 01 Recording B formed a group (there were no channel or speaking style differences). C_{llr} was also calculated using these group means. See [15,19] for details of the procedures for calculating validity and reliability.

Table 1: Log-likelihood-ratio cost and the 95% credible interval in $\log_{10}$ (LR) for each system.

system	C_{llr}	95% CI
acoustic-phonetic	0.349	±2.83
automatic	0.029	±1.26
fused	0.009	±3.41

Tippett plots of the results from the three systems are provided in Figures 1-3 (solid lines show group means and dashed lines to the left and right indicate the 95% CI).

Figure 1: Tippett plot of test results from the acoustic-phonetic system.

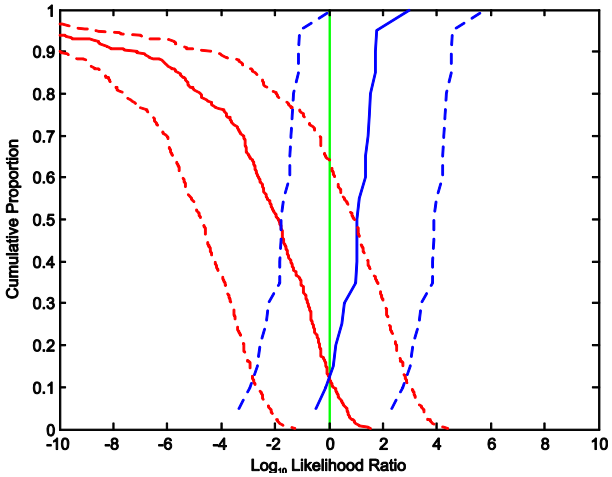


Figure 2: Tippett plot of test results from the automatic system.

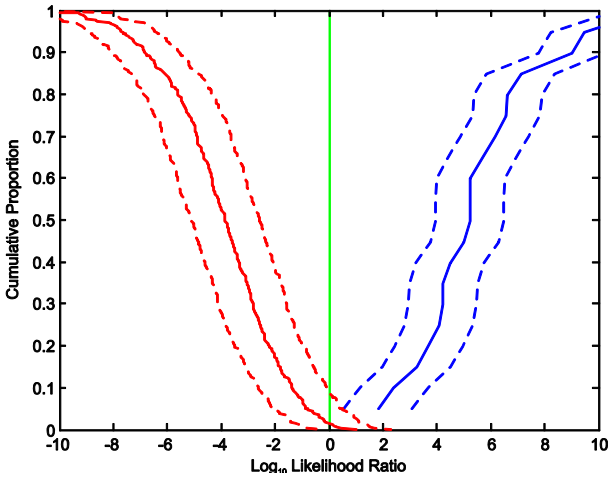
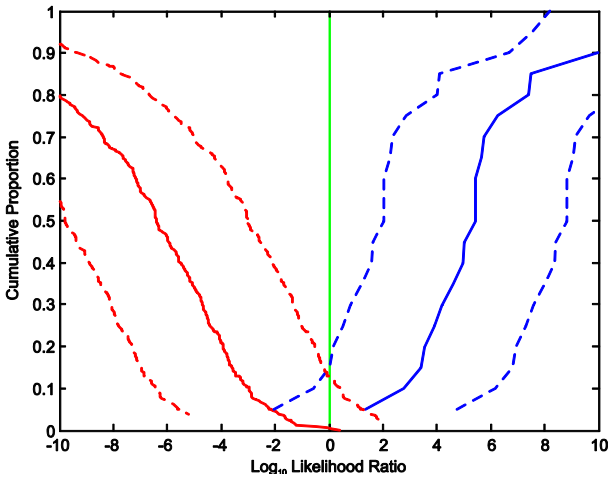


Figure 3: Tippett plot of test results from the fused system.



4. DISCUSSION AND CONCLUSION

With respect to validity, test results indicated that adding information extracted from the formant trajectories of Chinese /iau/ tokens lead to substantial improvement over a generic automatic forensic-voice-comparison system which did not explicitly exploit acoustic-phonetic information – C_{llr} of the fused system was approximately one third of that of the automatic system. With respect to reliability however, the 95% credible interval for the fused system was more than twice that of the automatic system. Given these results, it is not clear whether the additional human labor required for the acoustic-phonetic procedure is justified. The present study used high-quality audio recordings and the performance of the automatic system alone was very good. The effectiveness of adding the acoustic-phonetic information may be better tested on a more challenging test set. Additional research should investigate the application of this technique to more forensically realistic conditions, including speaking-style and recording/transmission-channel mismatch.

Future research could investigate whether aspects of the acoustic-phonetic procedure can be automated while still maintaining the level of validity and reliability obtained using the labor-intensive version. Research could also be conducted to explore whether information extracted from other phonetic units in addition to /iau/ can be combined to obtain greater degrees of validity and reliability (see [10, 14]).

An alternative to applying the MVKD formula to the acoustic-phonetic data would be to apply the GMM-UBM procedure. [14] found that GMM-UBM substantially outperformed MVKD in both validity and reliability when several acoustic-phonetic systems, each based on a different phonetic unit, were fused, (although performance on individual phonetic units was similar). Preliminary investigations of applying the GMM-UBM procedure to the data from the single phonetic unit in the current study found considerably worse performance than that obtained for the MVKD procedure. There was also instability in the modeling of the UBM, revealed by repeatedly training the UBM from scratch on the same training data but using different random seeds. The difference in the relative performance of MVKD and GMM-UBM between the present paper and [14] is probably due to the larger number of features being modeled and sparser

data: 10 coefficient values from two formant trajectories of a triphthong in the present paper with between 8 and 20 tokens per speaker per recording, compared to 4 coefficient values from a single formant trajectory of each of a number of diphthongs in [14] with between 16 and 20 tokens per speaker per recording. Compared to the GMM-UBM procedure, the MVKD procedure has a higher potential bias but lower potential variance, and the latter seems to handle the high dimensionality and sparse data of the present study better.

5. ACKNOWLEDGMENTS

Data collection was funded by an International Association of Forensic Phonetics and Acoustics (IAFPA) Research Grant. Analysis and paper writing were funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

6. REFERENCES

[1] Aitken, C.G.G., Lucy, D. 2004a. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* 53, 109-122. doi:10.1046/j.0035-9254.2003.05271.x

[2] Brümmer, N. 2005. Tools for fusion and calibration of automatic speaker detection systems. *http://niko.brummer.googlepages.com/focal*

[3] Brümmer, N., Burget, L., Cernocký, J.H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D.A., Matejka, P., Schwarz, P., Strasheim, A. 2007. Fusion of heterogenous speaker recognition systems in the STBU. *IEEE Trans. Audio, Speech, Lang. Process*. 15, 2072-2084. doi:10.1109/TASL.2007. 902870.

[4] Brümmer, N., du Preez, J. 2006. Application independent evaluation of speaker detection. *Comp. Speech Lang.* 20, 230–275. doi:10.1016/j.csl.2005.08.001

[5] Enzinger, E. 2010. Characterising formant tracks in Viennese diphthongs for forensic speaker comparison. *Proceedings of the 39th Audio Engineering Society Conference – Audio Forensics: Practices and Challenges* Hillerød, Denmark, 47-52.

[6] Furui, S. 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Proc. IEEE Trans. Acoust., Speech and Sig.* 34, 52-59. doi:10.1109/TASSP.1986.1164788

[7] van Leeuwen, D.A., Brümmer, N. 2007. An introduction to application-independent evaluation of speaker recognition systems. In Müller, C. (ed.), *Speaker Classification I: Fundamentals, Features, and Methods*. Heidelberg, Germany: Springer-Verlag, 330-353. doi:10.1007/978-3-540-74200-5_1

[8] Morrison, G.S. 2007. Multivar_kernel_LR: Matlab implementation of Aitken & Lucy's (2004), Forensic likelihood-ratio software using multivariate-kernel-density estimation. *http://geoff-morrison.net/*

[9] Morrison, G.S. 2009. Forensic voice comparison and the paradigm shift. *Sci. & Justice* 49, 298-308. doi:10.1016/j.scijus.2009.09.002

[10] 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. doi:10.1121/1.3081384

[11] Morrison, G.S. 2009. Robust version of train_llr_fusion.m from Niko Brümmer's FoCal Toolbox, release 2009-07-02. *http://geoff-morrison.net/*

[12] Morrison, G.S. 2010. Forensic voice comparison. In Freckelton, I., Selby, H. (eds.), *Expert Evidence*. Sydney, Australia: Thomson Reuters.

[13] Morrison, G.S. 2010. SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files. Release 2010-11-18. *http://geoff-morrison.net/*

[14] Morrison, G.S. 2011. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM). *Speech Commun.* 53, 242-256. doi:10.1016/j.specom.2010.09.005

[15] Morrison, G.S. 2011. Measuring the validity and reliability of forensic likelihood-ratio systems. *Sci. & Justice*, published online 14 April 2011. doi:10.1016/j.scijus.2011.03.002

[16] Morrison, G.S., Nearey, T.M. 2010. Formant Measurer: Software for efficient human-supervised measurement of format trajectories, release 2010-11-30. *http://geoff-morrison.net/*

[17] Morrison, G.S., Rose, P., Zhang, C., submitted 2011. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice.

[18] Morrison, G.S., Thiruvaran, T., Epps, J. 2010. An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* Melbourne, Australia, 74-77.

[19] Morrison, G.S., Thiruvaran, T., Epps, J., 2010. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. *Proceedings of Odyssey 2010, The Speaker and Language Recognition Workshop* Brno, Czech Republic, 63-70.

[20] Morrison, G.S., Zhang, C., Rose, P. 2010. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Sci. Int.* 208, 59-65. doi:10.1016/j.forsciint.2010.11.001

[21] Pigeon, S., Druyts, P., Verlinde, P. 2000. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digit. Signal Process.* 10, 237-248. doi:10.1006/dspr.1999.0358

[22] Reynolds, D.A., Quatieri, T.F., Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* 10, 19-41. doi:10.1006/dspr.1999.0361

[23] Zhang, C., Morrison, G.S. 2011. Forensic database of audio recordings of 68 female speakers of Standard Chinese. *http://databases.forensic-voice-comparison.net/*