Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1pSCc: Distinguishing Between Science and Pseudoscience in Forensic Acoustics II

1pSCc6. Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/

Cuiling Zhang* and Ewald Enzinger

***Corresponding author's address: Department of Foresnic Science & Technology, China Criminal Police Universty, 83 Tawan Street, Shenyang, 110854, Liaoning, China, cuiling-zhang@forensic-voice-comparison.net**

This study investigates the fusion of multiple formant-trajectory- and fundamental-frequency-trajectory-based (f0-trajectory-based) forensic-voice-comparison systems. Each system was based on tokens of a single phoneme: tokens of Chinese /ei1/, /ai2/, and /iau1/ (numbers indicate tones). Human-supervised formant-trajectory and f0-trajectory measurements were made on tokens from a database of recordings of 60 female speakers of Chinese. Discrete cosine transforms (DCT) were fitted to the trajectories and the DCT coefficients used to calculate likelihood ratios via the multivariate kernel density (MVKD) formula. The individual-phoneme systems were fused with each other and with a baseline mel-frequency cepstral-coefficient (MFCC) Gaussian-mixture-model universal-background-model (GMM-UBM). The latter made use of the entire speech-active portion of the recordings. Tests were conducted using high-quality recordings as nominal suspect samples and mobile-to-landline transmitted recordings as nominal offender samples. Fusion of the phoneme-systems with the baseline system via logistic regression did not lead to any substantial improvement in validity, and reliability deteriorated.

Published by the Acoustical Society of America through the American Institute of Physics

INTRODUCTION

Recent research on acoustic-phonetic inspired forensic-voice-comparison systems has explored a number of individual aspects. Morrison (2009a) found that fusion of multiple systems based on formant-trajectory measurements of five Australian English diphthongs showed promising performance. Zhang *et al.* (2011) fused a formant-trajectory system based on Chinese /iau^1/ with a baseline automatic Mel-frequency-cepstral-coefficient (MFCC) based Gaussian-mixture-model–universal-background-model (GMM-UBM) system and compared the resulting performance with that of the baseline system. Recent studies explored both formant trajectories and fundamental-frequency (f0) trajectories of Cantonese /ɔy^2/ (Li and Rose, 2012) and /i^2/ (Wang and Rose, 2012).

All of the aforementioned studies used high-quality recordings. However, recordings provided for forensic-voice-comparison analysis are often collected under degraded conditions due to transmission channel or noise. Work on the effect of different telephone-transmission systems on formant-trajectory-based systems has shown that if there is a mobile-telephone transmission involved, fusion of a formant-trajectory-based system based on Chinese /iau^1/ with a baseline MFCC-based GMM-UBM system resulted in very little improvement in validity over the baseline system (Zhang *et al.*, 2012).

This study investigates all these different aspects in combination. Formant trajectories and f0 trajectories are made of tokens of Standard Chinese /ei^1/, /ai^2/, and /iau^1/. For each of the three phonemes an acoustic-phonetic systems is created that jointly models formant and f0 trajectories. Consequently, one or multiple of these three systems are fused with a baseline MFCC-based GMM-UBM system, and the performance is assessed relative to the baseline system. Experiments are made in a mismatched condition between high-quality suspect recordings and mobile-to-landline transmitted offender recordings.

METHODOLOGY

Data

The data were extracted from a database of two non-contemporaneous voice recordings of each of 60 female speakers of Standard Chinese (Zhang and Morrison, 2011). See Morrison *et al.* (2012) for details of the data collection protocol. 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 recieved 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. The first and second recording sessions were separated by 2-3 weeks. High-quality recordings were made at 44 100 samples per second 16 bit quantization using flat-frequency-response lapel microphones (Sennheiser MKE 2 P-C) and an external soundcard (Roland® UA-25 EX), with one speaker on each of the two recording channels.

In addition to the original high-quality recordings, degraded sets of recordings were created by passing the high-quality set of recordings through a mobile-to-landline transmission channel. The details of the procedure are described in Zhang *et al.* (2012). The high-quality condition was treated as the condition of the suspect (known identity) recording, and the mobile-to-landline condition was treated as the condition for the offender (questioned identity) recording.

Stressed tokens of /iau/ on tone 1, tokens of /ei/ on tone 1, and tokens of /ai/ on tone 2 were manually located and marked using SOUNDLABELLER (Morrison, 2010). Table 1 contains the number of tokens per segment. Unlike an earlier study (Zhang *et al.*, 2011) in which all the /iau/ tokens were the realizations of a single word ("yao" one) those in the present study were also taken from the obstruent-initial open-syllable contexts /piau/ "biao" and /tɕiau/ "jiao" (Standard Chinese contrasts voiceless plosives and affricates, as in these words, versus voiceless-aspirated plosives and affricates).

TABLE 1: Number of tokens of each segment per speaker per session.

Segment	Median	Min.	Max.
/ei^1/	22	5	44
/ai^2/	12	6	26
/iau^1/	21.5	6	41

Forensic-voice-comparison systems

Baseline MFCC GMM-UBM system

The baseline forensic-voice-comparison system extracted 16 mel-frequency-cepstral-coefficients (MFCCs) 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 (Furui, 1986). Gaussian short-term feature warping (Pelecanos and Sridharan, 2001) using a three second sliding window was applied to the MFCCs and deltas before subsequent modeling. A GMM-UBM model (Reynolds *et al.*, 2000) 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 set to 1024. Extraction of MFCCs and training of GMMs was performed using an implementation provided by the Hidden Markov Toolkit (Young, 1993).

Formant-trajectory- and f0-trajectory-based systems

Human-supervised measurements of the trajectories of the first three formants (F1, F2, and F3) and the fundamental frequency of each vowel token were made using FORMANTMEASURER (Morrison and Nearey, 2011). This software is based on the formant tracking procedure outlined in Nearey *et al.* (2002). Fundamental frequency tracks were measured using the autocorrelation algorithm of Boersma (1993). See Zhang *et al.* (2013, 2012) for details on the procedure for human-supervised formant-trajectory and fundamental-frequency-trajectory measurement.

Discrete cosine transforms (DCTs) were fitted to the measured formant and f0 trajectories of all the /ei^1/, /ai^2/, and /iau^1/ tokens. The number of formant DCT coefficients (zeroth through second, zeroth through third, or zeroth through fourth) was selected on the basis of tests made on the development set. In the same way it was decided whether to use f0 trajectories and, if so, whether to include only the zeroth or both the zeroth and the first DCT coefficient. See Table 2 for the DCT coefficients used in each system. Likelihood ratios were calculated using the multivariate kernel density (MVKD) formula (Aitken and Lucy, 2004, implemented in Morrison, 2007).

TABLE 2: DCT coefficients selected to represent formant-frequency and fundamental-frequency trajectories.

Segment	formants	f0
/ei^1/	DCT0–DCT4	not included
/ai^2/	DCT0–DCT2	DCT0, DCT1
/iau^1/	DCT0–DCT3	DCT0, DCT1

Use of background, development, and test sets

In the tests of forensic-voice-comparison systems described below, tokens from the first 20 speakers were used as background data, data from the next 20 speakers were used as development data, and data from the last 20 speakers were used as test data.

In both the development and test sets, every speaker's Session 1 recording (nominal offender recording) was compared with their own Session 2 recording (nominal suspect recording) for a same-speaker comparison and with every other speaker's Session 1 as well as Session 2 recording (nominal suspect recordings) as different-speaker comparisons. The nominal offender recordings were mobile-to-landline transmitted recordings, and the nominal suspect recordings and the background were high-quality recordings. Both Session 1 and Session 2 recordings were included in the background. The development set was used to calculate scores which were then used to calculate weights for logistic-regression calibration (Brümmer and du Preez, 2006; van Leeuwen and Brümmer, 2007; Morrison, 2013) which was applied to convert the scores from the test set to likelihood ratios (calculations were performed using Brümmer, 2005, and Morrison, 2009b). Logistic regression was also used to fuse the scores from multiple individual systems and convert them to likelihood ratios (Pigeon *et al.*, 2000).

RESULTS

The validity and reliability of the systems was evaluated using the log likelihood-ratio cost (C_{llr}) as a metric of validity (accuracy), and an estimate of the 95% credible interval (95% CI) as a metric of reliability (precision) (Morrison, 2011; Morrison *et al.*, 2010) (C_{llr} was calculated using the mean procedure and the 95%CI using the parametric procedure).

Figure 1 gives an overview of the results. The baseline system had a C_{llr} of 0.152 and a $\log_{10}$ 95% CI of 1.50 (a Tippett plot is provided in the left panel of Figure 2). Of the fusions of individual acoustic-phonetic systems with the baseline system, no fused system outperformed the baseline system in both validity and reliability.

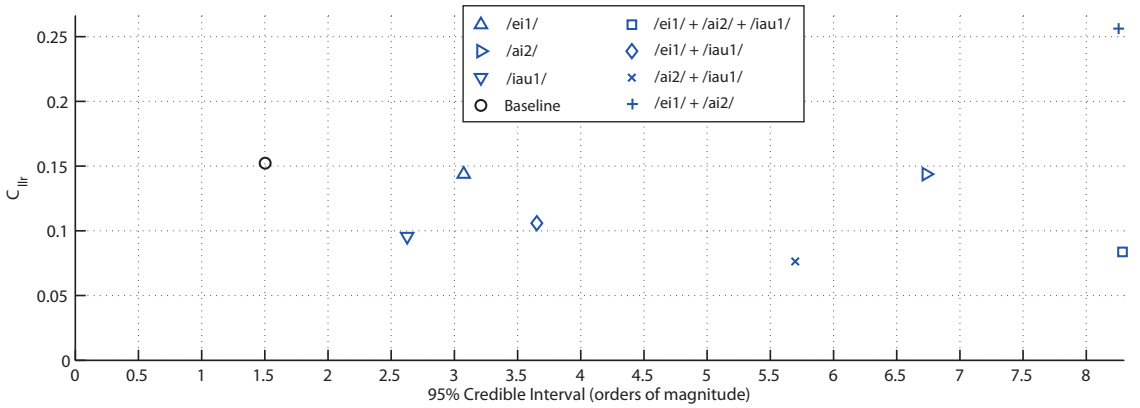

FIGURE 1: Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for fusions of one or multiple formant-trajectory- and f0-trajectory-based systems with the baseline MFCC-based GMM-UBM system (mobile-to-landline v high-quality recordings).

Fusion of the formant-trajectory- and f0-trajectory-based system on /iau[1]/ with the baseline system provides substantial increase in validity (C_{llr} 0.0956, −37%), but at a large decrease in reliability ($\log_{10}$ 95% CI 2.63, +75%). Comparison of the Tippett plots in Figure 2 indicates that the improvement in C_{llr} for the formant-trajectory- and f0-trajectory-based system was mostly due to already large magnitude negative log likelihood ratios from different-speaker comparisons getting even larger, rather than due to a reduction in the number and magnitude of log likelihood ratios which gave greater support to contrary-to-fact rather than consistent-with fact hypotheses. At the same time, the 95% credible interval increased by more than an order of magnitude. Fusion of one or multiple other systems with the baseline system resulted in even more severe deterioration of reliability.

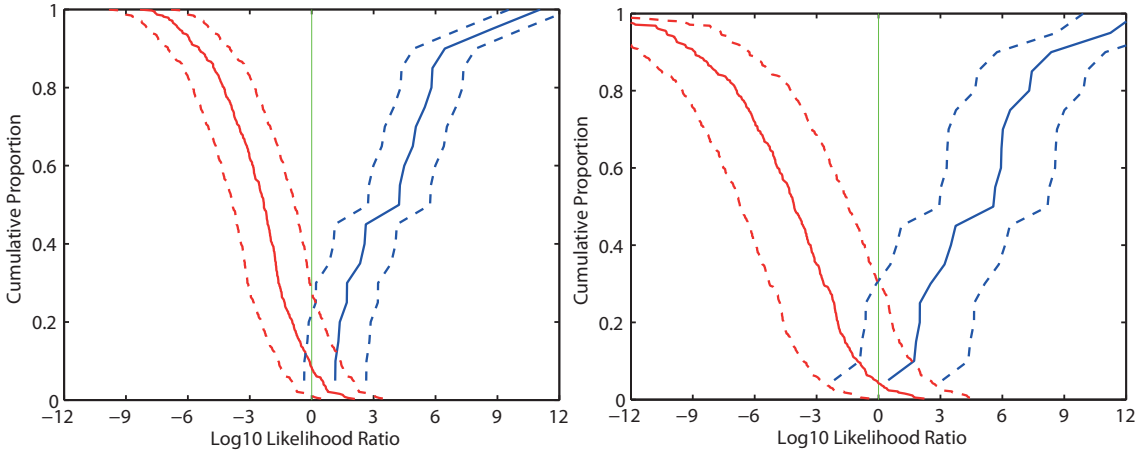

FIGURE 2: Tippett plot of the baseline MFCC-based GMM-UBM system (left) and after fusion with the formant-trajectory and f0-trajectory-based system on /iau1/ (right) (mobile-to-landline v high-quality recordings).

DISCUSSION AND CONCLUSIONS

The fusion of Chinese /ei^1/, /ai^2/, and /iau^1/ based formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems with a baseline system was assessed. None of the fused systems outperformed the baseline systems in both validity and reliability. While substantial improvements in validity were observed for the system based on /iau^1/ after fusion with the baseline, reliability deteriorated. These results are consistent with those found in a previous study on /iau^1/ on high-quality recordings (Zhang *et al.*, 2011).

Earlier studies have also found that reliability deteriorates when combining multiple systems using logistic-regression fusion (Enzinger *et al.*, 2012; Enzinger and Morrison, 2012). Further research is needed to investigate alternative methods for combining different systems.

ACKNOWLEDGMENTS

This research received support from the following sources:

- The Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142.
- The China Scholarship Council State-Sponsored Scholarship Program for Visiting Scholars.
- The Ministry of Education of the People's Republic of China "Program for New Century Excellent Talents in University" (NCET-11-0836).
- An International Association of Forensic Phonetics and Acoustics Research Grant.

Unless otherwise explicitly attributed, the opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations.

REFERENCES

Aitken, C. G. G. and Lucy, D. (**2004**). "Evaluation of trace evidence in the form of multivariate data", Applied Statistics, **53**, 109–122.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", IFA Proceedings, **17**, 97–110.

Brümmer, N. (**2005**). "Tools for fusion and calibration of automatic speaker detection systems", URL `http://niko.brummer.googlepages.com/focal`.

Brümmer, N. and du Preez, J. (**2006**). "Application-independent evaluation of speaker detection", Computer Speech and Language, **20**, 230–275.

Enzinger, E. and Morrison, G. S. (**2012**). "The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems", in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST-2012)* (Sydney, Australia).

Enzinger, E., Zhang, C., and Morrison, G. S. (**2012**). "Voice source features for forensic voice comparison – an evaluation of the glottex software package", in *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop*, 78–85. (Singapore).

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

Li, J. and Rose, P. (**2012**). "Likelihood Ratio-Based Forensic Voice Comparison with F-Pattern and Tonal F0 from the Cantonese /ɔy/ Diphthong", in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012)*, 201–204 (Sydney, Australia).

Morrison, G. S. (**2007**). "multivar_kernel_LR: Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multi-variate kernel density estimation", URL `http://geoff-morrison.net/#MVKD`.

Morrison, G. S. (**2009**a). "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs", J. Acoust. Soc. Am., **125**, 2387–2397, doi:10.1121/1.3081384.

Morrison, G. S. (**2009**b). "Robust version of train_llr_fusion.m from Niko Brümmer's FoCaL Toolbox (release 2009-07-02)", URL `http://geoff-morrison.net/#TrainFus`.

Morrison, G. S. (**2010**). "Soundlabeller: Ergonomically designed software for marking and labelling portions of sound files (release 2010-11-18)", URL `http://geoff-morrison.net/#SndLbl`.

Morrison, G. S. (**2011**). “Measuring the validity and reliability of forensic likelihood-ratio systems”, Science & Justice, **51**, 91–98, doi:10.1016/j.scijus.2011.03.002.

Morrison, G. S. (**2013**). “Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio”, Australian Journal of Forensic Sciences, doi:10.1080/00450618.2012.733025.

Morrison, G. S. and Nearey, T. M. (**2011**). “FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories”, URL `http://geoff-morrison.net/#FrmMes`.

Morrison, G. S., Rose, P., and Zhang, C. (**2012**). “Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice”, Australian Journal of Forensic Sciences, **44**, 155–167, doi:10.1080/00450618.2011.630412.

Morrison, G. S., Thiruvaran, T., and Epps, J. (**2010**). “Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system”, in *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop*, edited by H. Cernocký and L. Burget, 63–70 (International Speech Communication Association, Brno, Czech Republic).

Nearey, T. M., Assmann, P. F., and Hillenbrand, J. M. (**2002**). “Evaluation of a strategy for automatic formant tracking”, J. Acoust. Soc. Amer., **112**, 2323.

Pelecanos, J. and Sridharan, S. (**2001**). “Feature warping for robust speaker verification”, in *Proceedings of the Odyssey Speaker Recognition Workshop* (International Speech Communication Association).

Pigeon, S., Druyts, P., and Verlinde, P. (**2000**). “Applying Logistic Regression to the Fusion of the NIST’99 1-Speaker Submissions”, Digital Signal Process., **10**, 237–248, doi:10.1006/dspr.1999.0358.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (**2000**). “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Process., **10**, 19–41.

van Leeuwen, D. A. and Brümmer, N. (**2007**). “An introduction to application-independent evaluation of speaker recognition systems”, in *Speaker Classification I. Fundamentals, Features, and Methods*, edited by C. Müller, 330–353 (Springer).

Wang, C. Y. and Rose, P. (**2012**). “Likelihood Ratio-Based Forensic Voice Comparison with Cantonese /i/ F-Pattern and Tonal F0”, in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012)*, 209–212 (Sydney, Australia).

Young, S. (**1993**). “The HTK hidden Markov model toolkit: Design and philosophy”, Technical Report, Department of Engineering, Cambridge University, U.K.

Zhang, C. and Morrison, G. S. (**2011**). “Forensic database of audio recordings of 68 female speakers of standard chinese”, URL `http://databases.forensic-voice-comparison.net/`.

Zhang, C., Morrison, G. S., Enzinger, E., and Ochoa, F. (**2012**). “Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. laboratory report, forensic voice comparison laboratory, school of electrical engineering & telecommunications, university of new south wales”, Technical Report.

Zhang, C., Morrison, G. S., Ochoa, F., and Enzinger, E. (**2013**). “Reliability of human-supervised formant-trajectory measurement for forensic voice comparison”, J. Acoust. Soc. Amer., **133**, EL54–EL60, doi:10.1121/1.4773223.

Zhang, C., Morrison, G. S., and Thiruvaran, T. (**2011**). “Forensic voice comparison using Chinese /iau/”, in *Proceedings of the 17th International Congress of Phonetic Sciences*, 2280–2283 (Hong Kong, China).