Testing the validity and reliability of forensic voice comparison based on reassigned time-frequency representations of Chinese /iau/

Ewald Enzinger[1,2]

[1]*Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia*

[2]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, 1040, Austria*

e.enzinger@student.unsw.edu.au

***Abstract*—Time-frequency reassigned representations of vowel segments have repeatedly been proposed as features for performing forensic voice comparison (FVC). Combined with a thresholding on the second-order mixed partial derivative of the short-time Fourier transform (STFT) phase to reduce artifacts and noise (pruning) such representations offer a sharpened representation of AM/FM components in the speech signal. We test the validity and reliability of FVC systems based on pruned reassigned time-frequency representations extracted from manually labeled /iau/ segments from a database of voice recordings of 60 female Chinese speakers. Logistic-regression fusion was used to combine the systems with a baseline MFCC GMM-UBM system. Performance was assessed as degree of improvement over the baseline system. Three recording channel conditions were tested: high-quality v high-quality, mobile-to-landline v mobile-to-landline, and mobile-to-landline v high-quality. For the latter two conditions improvements over the baseline system were observed in terms of validity, but with decreases in reliability.**

I. Introduction

Time-frequency reassignment (*TFR*) of the short-term Fourier transform (STFT) offers a sharpened representation of impulses in the signal as well as instantaneous frequencies of line components. The coefficients obtained in the STFT, $X_h(\omega, T)$, of a speech signal are spaced on the time-frequency grid based on the length and overlap of the moving window $h(t)$. The *channelized instantaneous frequency* (CIF) and *local group delay* (LGD) for each time-frequency bin are defined as [1], [2]:

$$\mathrm{CIF}(\omega, T) = \frac{\delta}{\delta T} \arg(X_h(\omega, T)) \quad (1)$$

$$\mathrm{LGD}(\omega, T) = \frac{\delta}{\delta \omega} \arg(X_h(\omega, T)) \quad (2)$$

In the reassignment the STFT coefficients are moved from the time and frequency center of the window in the STFT to the center of gravity of the energy inside the window specified by the CIF and the time shifted by the LGD. After time-frequency reassignment the amplitudes can be thresholded to reduce artifacts and noise using a *pruning* method [3] based on the second-order mixed partial derivative of the STFT phase, originally proposed by [4]. Dependent on the thresholding criterion, quasi-sinusoidal components or impulses are retained, thereby emphasizing vocal tract resonances or glottal pulsations.

Following interest in the use of time-frequency reassignment methods for speech analysis (e.g. [1], [3], [5]), recent publications proposed its use in forensic voice comparison (FVC). Fulop and Disner [6], [7] suggest lower within- than between-speaker variability in purely visual comparison of pruned reassigned spectrograms obtained from short utterances of speech with comparable voice quality. Recent presentation of the method at the 165th meeting of the Acoustical Society of America in Montréal [8] raised some concern among forensic speech scientists due to reference by the authors to the spectrographic approach as well as the use of the term 'voiceprint'. This approach is based on a human expert making decisions on this basis of visual inspection of spectrograms, and has a controversial history going back to the 1960s (See [9]–[12] for reviews of the scientific and legal debate); however, the output of the procedure for creating pruned reassigned spectrograms can also be used as the basis for quantitative measurements which can then be used to train and test statistical models. An automatic classification experiment based on quantitative measurements instead of visual comparison presented in [8] suggested that the approach had some potential but was based on only six speakers.

In the present paper we assess the performance of likelihood-ratio forensic-voice-comparison systems based on pruned reassigned time-frequency representations. Two feature extraction approaches, one proposed in [8] and a novel approach using the two-dimensional discrete cosine transform, were evaluated using manually labelled tokens of /iau/ taken from a database of recordings of 60 female Chinese speakers. Logistic-regression fusion was used to combine the system with a baseline mel frequency cepstral coefficient (MFCC) Gaussian mixture model - universal background model (GMM-UBM) system [13]. Tests were made in three different channel conditions: high-quality v high-quality, mobile-to-landline v mobile-to-landline, and mobile-to-landline v high-quality. Performance was assessed as degree

WIFS'2013, November 18-21, 2013, Guangzhou, China.
ISBN 978-1-4673-5593-3

of improvement over the baseline system[1].

II. Methodology

A. Data base

The evaluation is based on a database of two non-contemporaneous voice recordings of each of 60 female speakers of Standard Chinese (Mandarin/Putonghua) [18]. See [19] 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. 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 [14]. System performance was assessed in three channel conditions:

- high-quality v high-quality
- mobile-to-landline v mobile-to-landline
- mobile-to-landline v high-quality

The condition to the right was treated as the condition of the suspect (known identity) recording, and the condition to the left was treated as the condition for the offender (questioned identity) recording.

Stressed tokens of /iau/ on tone 1 ("yao" one) were manually located and marked using SoundLabeller [20]. The number of tokens per recording ranged between 8 and 20.

B. Forensic-voice-comparison systems

1) 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 [21]. Feature warping [22] using a three second sliding window was applied to the MFCCs and deltas before subsequent modeling. A GMM-UBM model [13] 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 [23].

[1]This paper closely follows other papers we have written on extracting information from acoustic signals ([14]–[17]). We have replicated the description of procedures used in earlier papers, in particular the data base (Sections II-A and II-C) and the baseline system (Sections II-B1 and II-B3).

2) Time-frequency reassignment based systems: The evaluation of pruned reassigned time-frequency representations (*TFR*) based features follows closely the feature extraction procedures proposed in [8]. First the short-time Fourier transform (STFT) was obtained from manually labelled tokens of /iau/. A Kaiser window with $\alpha = 3$ was applied to the signal. The window length was set to approximately 90% of one glottal cycle estimated from the average fundamental frequency, which was determined under human supervision using the algorithm in [24]. A step size of 2 samples was used. Contrary to [8], where the vowel tokens were truncated to 40 ms, we computed the STFT over the whole segment duration.

Channelized instantaneous frequencies (CIF) and local group delay (LGD) were calculated using an approximate method based on the phase gradient of the STFT [1], [4]. The phase derivatives of the STFT both in frequency and in time were computed as the arguments of cross-spectral surfaces:

$$C(\omega, T, \epsilon) = X_h(\omega, T + \frac{\epsilon}{2}) X_h^*(\omega, T - \frac{\epsilon}{2}) \quad (3)$$

$$L(\omega, T, \epsilon) = X_h(\omega + \frac{\epsilon}{2}, T) X_h^*(\omega - \frac{\epsilon}{2}, T) \quad (4)$$

The CIF and LGD were then obtained as:

$$\mathrm{CIF}(\omega, T) \approx \frac{1}{\epsilon} \arg(C(\omega, T, \epsilon)) \quad (5)$$

$$\mathrm{LGD}(\omega, T) \approx \frac{1}{\epsilon} \arg(L(\omega, T, \epsilon)) \quad (6)$$

The algorithmic details are described in [1] (we used a MATLAB® implementation provided by the authors[2]). The time and frequency corrected STFT magnitudes were then pruned based on the second-order mixed partial derivative of the STFT phase [3]. The pruning thresholds for the mixed partial derivative were set to < 0.1 to retain line components and to between 0.75 and 1.25 to retain impulses [7]. The normalized magnitudes were further thresholded at -80 dB. Figure 1 shows an example of a pruned time-frequency-reassigned spectrogram of a token of /iau/ in the database.

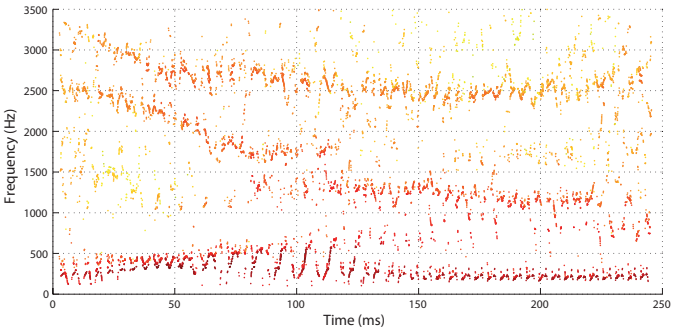

Fig. 1. Example of a pruned reassigned spectrogram of /iau/.

In this paper we tested two approaches for obtaining low-dimensional feature representations from pruned time-frequency reassigned STFT magnitudes. The first approach

[2]Available at http://seanfulop.weebly.com/research.html

was originally proposed in [8]. The time-frequency reassigned STFT magnitudes were discretized in time and frequency using a coarse grid with 50 time and 85 frequency bins. The bins were linearly spaced over time and frequencies between 100 and 3500 Hz. From this coarse grid the average was calculated over the bins to obtain time- and frequency-averaged features (*TFR AVG*). Figure 2 shows the discretized grid of the example in Figure 1 as well as the time- and frequency-averaged features. Principal component analysis was used to

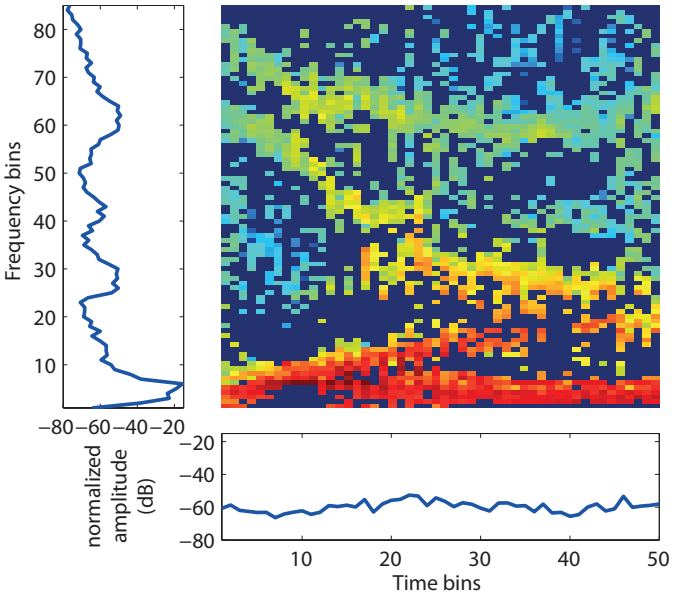

Fig. 2. Discretized pruned time-frequency reassigned distributions of the example spectrogram given in Figure 1, along with the time- and frequency-averaged features.

further reduce the dimensionality, retaining 10 time features and 20 frequency features (The number of components was selected empirically). We pooled all features extracted from speakers in the background set (see Section II-C) and used them to obtain a projection matrix from the eigenvectors of the covariance matrix. The concatenated time- and frequency-averaged feature vector had a dimension of 30.

The above characterization of the time-frequency representation is theoretically inadequate for speech segments that have significant correlation over time and frequency, such as the non-linear correlation in the triphthong /iau/. We therefore propose a second approach based on the two-dimensional discrete cosine transform (*TFR DCT*). First, the time-frequency reassigned distributions were again discretized in time and frequency. We used 85 frequency bins and set the number of time bins to be equal to the segment duration in milliseconds. Then, we computed the 2D DCT, from which we retained the lower-order 7×7 coefficients (The number of coefficients was experimentally determined in tests on the development set using 4×4, 5×5, 6×6, and 7×7 coefficients). The coefficients were scaled according to the number of time (N) and frequency (M) bins in the grid using a factor of $(1/\sqrt{NM})$. After discarding the zeroth coefficient, we used the remaining 48 coefficients as features in an FVC system. Figure 3 shows the discretized grid of the example in Figure 1 as well as after reconstruction from the lower-order 7×7 coefficient values by applying the inverse 2D DCT.

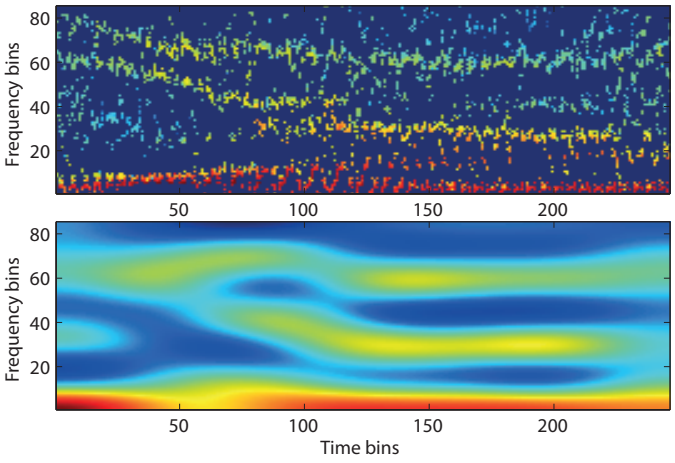

Fig. 3. Discretized representation of the example pruned reassigned spectrogram in Figure 1, as well as after reconstruction from the lower-order 7×7 coefficient values by applying the inverse 2D DCT.

Two FVC systems were created based on each of the sets of features. In both systems, extracted feature vectors were modeled using the GMM-UBM approach [13]. A UBM with diagonal covariance matrices was trained using feature vectors from the set of background speakers pooled across both recording sessions. Suspect GMMs were trained via maximum a-posteriori (MAP) adaptation from the UBM. The optimal number of Gaussians in the mixture (2, 4, 8, 16, 32, 64) were empirically determined via tests using the development set. A score was calculated as

$$\text{score} = \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{p(x_i|\lambda_{suspect})}{p(x_i|\lambda_{UBM})} \right), \qquad (7)$$

where x_i is a feature vector, N is the number of tokens of /iau/ in the offender data, and $\lambda_{suspect}$ and λ_{UBM} represent the models of the suspect and the background, respectively. In the following the FVC system based on time- and frequency averages is abbreviated as *TFR AVG* and the FVC system based on 2D DCT coefficients as *TFR DCT*.

3) MFCC-on-/iau/ system: A second MFCC-based system was constructed which was identical to the baseline system except that MFCCs and deltas were only calculated for the portions of the recordings which fell within the /iau/ markers. This system uses the same portions of the recordings as the systems based on time-frequency reassignment and is thus a diagnostic as to whether it is the selection of the /iau/ tokens which is important or whether the time-frequency features also contribute to system performance.

C. Use of background, development, and test sets

In the tests of FVC systems described below, data 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. 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 [25]–[27] which was applied to convert the scores from the test set to likelihood ratios (calculations were performed using [28], and [29]). Logistic regression was also used to fuse the scores from multiple individual systems and convert them to likelihood ratios [30].

III. 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) [31], [32] (C_{llr} was calculated using the mean procedure [31, §3.3] and the 95%CI using the parametric procedure [32, §2.3]). Readers familiar with automatic speaker recognition but not forensic voice comparison should note that metrics such as equal error rate (EER) and plots such as detection error trade-off (DET) [33] are not presented here since they are based on imposing hard thresholds on posterior probabilities and are therefore incompatible with the likelihood-ratio framework for the evaluation of forensic evidence [31], [34].

A. High-quality v high-quality recordings

Figure 4 shows the results of systems based on time- and frequency averages (*TFR AVG*) and 2D DCTs (*TFR DCT*) of reassigned time-frequency representations, and systems based on MFCCs for the high-quality v high-quality tests. The baseline system had a C_{llr} of 0.019 and a $\log_{10}$ 95% CI of 1.51. Of the fusions of both time-frequency-reassignment based and the MFCC based systems with the baseline system, no fused system outperformed the baseline system in both C_{llr} and 95% CI.

B. Mobile-to-landline v mobile-to-landline recordings

Figure 5 shows the results for the mobile-to-landline v mobile-to-landline tests. The baseline system had a C_{llr} of 0.219 and a $\log_{10}$ 95% CI of 0.97. Fusion of the system based on 2D DCT coefficients of reassigned time-frequency representations (*TFR DCT*) showed some improvement in validity (C_{llr} 0.192, -12.3%), but at a decrease in reliability ($\log_{10}$ 95% CI 1.071, $+10.9\%$). Fusion of the MFCC-based system with the baseline system resulted in nearly identical values.

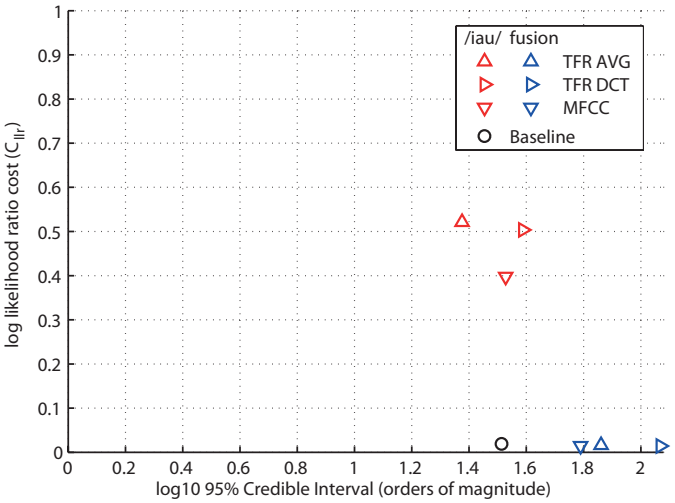

Fig. 4. Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for systems based on time- and frequency averages (*TFR AVG*) and 2D DCTs (*TFR DCT*) of reassigned time-frequency representations, and MFCCs individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (high-quality v high-quality recordings).

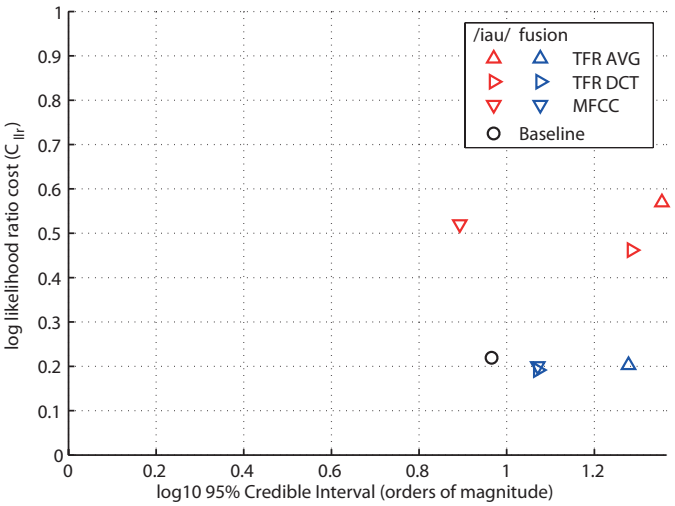

Fig. 5. Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for systems based on time- and frequency averages (*TFR AVG*) and 2D DCTs (*TFR DCT*) of reassigned time-frequency representations, and MFCCs individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (mobile-to-landline v mobile-to-landline recordings).

C. Mobile-to-landline v high-quality recordings

Figure 6 shows the results for the mobile-to-landline v high-quality tests. The baseline system had a C_{llr} of 0.152 and a $\log_{10}$ 95% CI of 1.50. Fusion of the system based on 2D DCT coefficients of reassigned time-frequency representations (*TFR DCT*) resulted in substantial improvement in validity (C_{llr} 0.120, -26.3%) at a loss in reliability ($\log_{10}$ 95% CI 1.071, $+13.3\%$). Fusion of the *MFCC*-based system with the baseline system showed similar improvement in C_{llr} with less increase in the $\log_{10}$ 95% CI, resulting in slightly better performance than the fused *TFR DCT* system. Fusion of the system based

on time- and frequency averages (*TFR AVG*) did not improve upon the baseline. Examination of Tippett plots of the baseline

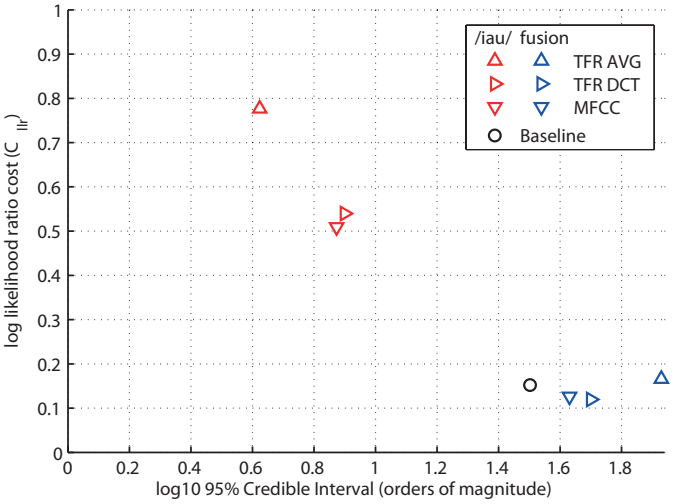

Fig. 6. Measures for validity (C_{llr}) and reliability ($log10$ 95% credible interval) for systems based on time- and frequency averages (*TFR AVG*) and 2D DCTs (*TFR DCT*) of reassigned time-frequency representations, and MFCCs individually (*red*) as well as after fusion with the generic fully-automatic baseline system (*blue*) (mobile-to-landline v high-quality recordings).

system and after fusion with the system based on 2D DCTs (*TFR DCT*) of reassigned time-frequency representations in Figure 7 indicates that reduction in C_{llr} was primarily due to large magnitude log likelihood ratios supporting consistent-with-fact hypotheses getting even larger, while the proportion of positive log likelihood ratios from different-speaker comparisons which contrary-to-fact gave greater support to the same-speaker hypothesis than to the different-speaker hypothesis was not reduced.

IV. Discussion and conclusion

The present paper assessed the performance of forensic voice comparison systems based on pruned reassigned time-frequency representations proposed by [6], [7]. Two approaches for feature extraction, time- and frequency averages proposed in [8] (*TFR AVG*) and using the 2D discrete cosine transform (*TFR DCT*) were computed from pruned reassigned time-frequency representations of tokens of /iau/ in a database of recordings of 60 female speakers of Standard Chinese. In both mobile-to-landline v high-quality and mobile-to-landline v mobile-to-landline conditions substantial relative improvements in validity were observed for the TFR DCT system after fusion with the baseline system, while reliability deteriorated. However, fusion of the MFCC-on-/iau/ system with the baseline system showed similar or slightly better performance than fusion of the TFR systems with the baseline system, indicating that the features based on time-frequency reassigned representations per se did not improve performance.

The approach for feature extraction based on the two-dimensional discrete cosine transform showed better performance than the original approach proposed in [8], which can be explained by the substantial non-linear correlation present

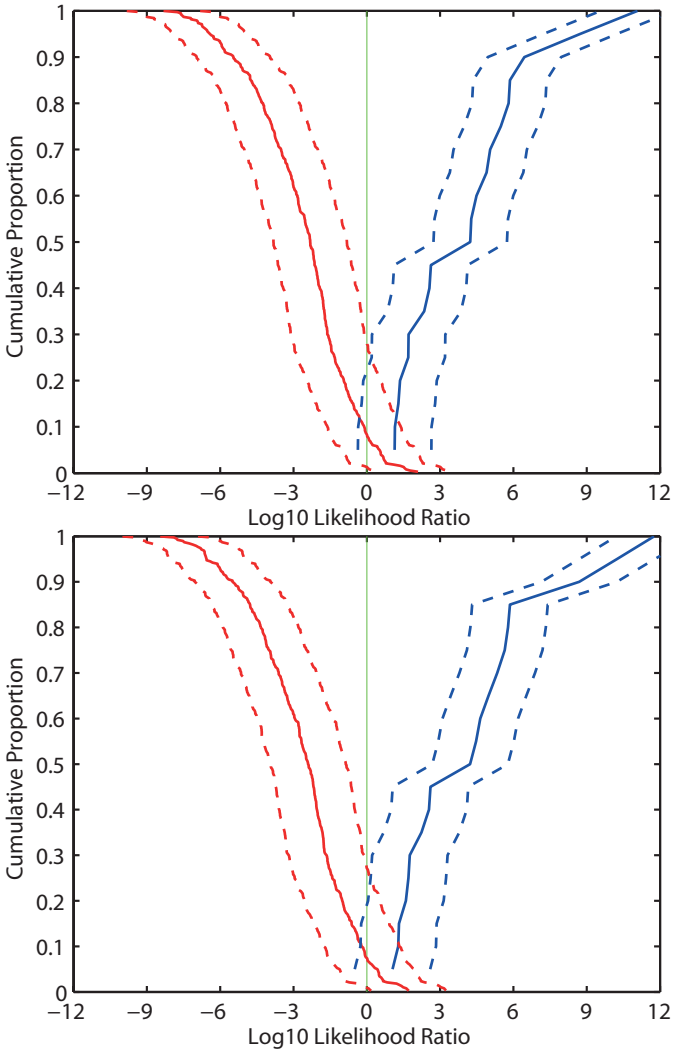

Fig. 7. Tippett plot of the baseline fully-automatic MFCC-based system (top) and after fusion with the system based on 2D DCTs (*TFR DCT*) of reassigned time-frequency representations (bottom) (mobile-to-landline v high-quality recordings).

in the time-frequency distribution of /iau/ triphthongs. By averaging over time and frequency bins this correlation is not taken into account. Thus, useful speaker-specific information is discarded.

However, the following caveats regarding these findings should be borne in mind: In [8] the authors used tokens of the vowel /æ/ from speakers of English, while this study uses the Chinese triphthong /iau/. We only tested recordings of female speakers using one speaking style. In forensic casework the signals (both suspect and offender) are also typically degraded by noise, reverberation, etc., and the effects of such additional signal degradation were not tested. While these findings can be seen as indication of performance under conditions similar to the ones tested, we consider testing of validity and reliability under conditions reflecting those of the case under investigation using data drawn from the relevant population as an essential principle for acceptable practice in

forensic voice comparison.

ACKNOWLEDGMENT

This research received support from 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.

Thanks to Geoffrey Stewart Morrison (Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales) and Cuiling Zhang (Department of Forensic Science & Technology, China Criminal Police University).

Opinions expressed are those of the author and do not necessary represent the policies of any of the above mentioned organisations.

REFERENCES

[1] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 360–371, January 2006.

[2] S. W. Hainsworth and M. D. Macleod, "Time frequency reassignment: A review and analysis," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.459, 2003.

[3] S. A. Fulop and K. Fitz, "Separation of components from impulses in reassigned spectrograms," *J. Acoust. Soc. Am.*, vol. 121, no. 3, pp. 1510–1518, March 2007.

[4] D. J. Nelson, "Cross-spectral methods for processing speech," *J. Acoust. Soc. Am.*, vol. 110, pp. 2575–2592, 2001.

[5] F. Plainte and W. A. Ainsworth, "Formant tracking using reassigned spectrum," in *Proc. Eurospeech*, Madrid, Spain, 18-21 September 1995, pp. 741–744.

[6] S. A. Fulop and S. F. Disner, "The reassigned spectrogram as a tool for voice identification," in *Proc. ICPhS XVI*, Saarbrücken, Germany, 6-10 August 2007, pp. 1853–1856.

[7] ——, "Advanced time-frequency displays applied to forensic speaker identification," in *Proceedings of Meetings on Acoustics*, vol. 6, 2009, p. 060008.

[8] S. A. Fulop and Y. Kim, "Speaker identification made easy with pruned reassigned spectrograms," in *Proceedings of Meetings on Acoustics*, vol. 19, 2013, p. 055068.

[9] J. S. Gruber and F. T. Poza, *Voicegram identification evidence*. Lawyers Cooperative Pub., 1995.

[10] P. Rose, *Forensic Speaker Identification*. Taylor & Francis, 2002.

[11] G. S. Morrison, *Expert Evidence*. Sydney, Australia: Thomson Reuters, 2010, ch. Forensic voice comparison.

[12] ——, "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison," *Science & Justice*, 2013.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[14] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices," *Speech Communication*, vol. 55, pp. 796–813, 2013.

[15] C. Zhang, G. S. Morrison, F. Ochoa, and E. Enzinger, "Reliability of human-supervised formant-trajectory measurement for forensic voice comparison," *J. Acoust. Soc. Amer.*, vol. 133, pp. EL54–EL60, 2013.

[16] E. Enzinger, C. Zhang, and G. S. Morrison, "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*, Singapore, June 2012, pp. 78–85.

[17] C. Zhang and E. Enzinger, "Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/," in *Proceedings of Meetings on Acoustics*, vol. 19. Montréal, Canada: Acoustical Society of America, 2-7 June 2013, p. 060044.

[18] C. Zhang and G. S. Morrison. (2011) Forensic database of audio recordings of 68 female speakers of Standard Chinese. [Online]. Available: http://databases.forensic-voice-comparison.net/

[19] G. S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Australian Journal of Forensic Sciences*, vol. 44, no. 2, pp. 155–167, 2012.

[20] G. S. Morrison. (2010) SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files (release 2010-11-18). [Online]. Available: http://geoff-morrison.net/#SndLbl

[21] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. 34, pp. 52–59, 1986.

[22] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of 2001: A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 18-22 June 2001, pp. 213–218.

[23] S. J. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.152, 6 September 1994.

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

[25] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[26] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I. Fundamentals, Features, and Methods*, C. Müller, Ed. Springer, 2007, pp. 330–353.

[27] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, pp. 173–197, 2013.

[28] N. Brümmer. (2005) Tools for fusion and calibration of automatic speaker detection systems. [Online]. Available: http://niko.brummer.googlepages.com/focal

[29] G. S. Morrison. (2009) Robust version of train_llr_fusion.m from Niko Brümmer's FoCaL Toolbox (release 2009-07-02). [Online]. Available: http://geoff-morrison.net/#TrainFus

[30] S. Pigeon, P. Druyts, and P. Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Process.*, vol. 10, pp. 237–248, 2000.

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

[32] G. S. Morrison, T. Thiruvaran, and J. Epps, "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*, H. Cernocký and L. Burget, Eds. Brno, Czech Republic: International Speech Communication Association, 2010, pp. 63–70.

[33] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece, 1997, pp. 1895–1898.

[34] G. S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM-UBM)," *Speech Communication*, vol. 53, pp. 242–256, 2010.