Likelihood ratio calculation in acoustic-phonetic forensic voice comparison: Comparison of three statistical modelling approaches

Ewald Enzinger[1]

[1]School of Elec. Eng. & Telecom., Univ. of New South Wales, Sydney, Australia

ewald.enzinger@entn.at

Abstract

This study compares three statistical models used to calculate likelihood ratios in acoustic-phonetic forensic-voice-comparison systems: Multivariate kernel density, principal component analysis kernel density, and a multivariate normal model. The data were coefficient values obtained from discrete cosine transforms fitted to human-supervised formant-trajectory measurements of tokens of /iau/ from a database of recordings of 60 female speakers of Chinese. Tests were conducted using high-quality recordings as nominal suspect samples and mobile-to-landline transmitted recordings as nominal offender samples. Performance was assessed before and after fusion with a baseline automatic mel frequency cepstral coefficient Gaussian mixture model universal background model system. In addition, Monte Carlo simulations were used to compare the output of the statistical models to true likelihood-ratio values calculated on the basis of the distribution specified for a simulated population.

Index Terms: forensic voice comparison, likelihood ratio, formant trajectories, validity, reliability

1. Introduction

In forensic inference and statistics there is wide support for the position that the logically correct way for a forensic scientist to evaluate the strength of forensic evidence is using a likelihood ratio [1, 2, 3]. A likelihood ratio is the probability of the observed evidence if the prosecution hypothesis were true versus if the defense hypothesis were true [4, 5]. Acoustic-phonetic approaches to forensic voice comparison predominantly used the multivariate kernel density (MVKD [6]) model to calculate likelihood ratios [7, 8, 9, 10, 11, 12] (see also a review in [13]). Concerns about the model's robustness, in particular numerical stability issues with the model's implementation and the difficulty of kernel density estimation, in particular when using a large number of features, prompted the development of the principal component analysis kernel density likelihood ratio (PCAKLR [14]) model. The method uses PCA to obtain a decorrelating transformation matrix and computes the likelihood ratio as the product of univariate likelihood ratios of the projected features. The univariate likelihood ratios are computed using a modified kernel density model [15, p. 338]. Another alternative to the MVKD model is a multivariate normal (MVN) model [6, 16], which assumes normal distributions for both within and between-speaker variation. While these assumptions may be a poor fit for many types of acoustic-phonetic features (see e.g. Rose [17, §5.1]), the model's lower variance may result in overall better performance. Recent experiments on acoustic-phonetically inspired approaches based on features of nasal segments have explored its use for statistical modeling [18]. A previous study comparing the MVKD model with the Gaussian mixture model universal background model (GMM-UBM [19]) approach for modeling formant-trajectory features extracted from Australian English diphthongs found that the latter achieved better validity and reliability when several systems, each based on a different phonetic unit, were fused [20]; however, a later study found that, when applied to a single phonetic unit, MVKD outperformed the GMM-UBM approach [21]. This was attributed to the high feature dimensionality and the relatively lower number of tokens available for training.

This study compares MVKD, PCAKLR, and MVN models for the calculation of likelihood ratios in formant-trajectory-based forensic-voice-comparison systems. The data were coefficient values obtained from discrete cosine transforms fitted to human-supervised formant-trajectory measurements of tokens of /iau/ from a database of recordings of 60 female speakers of Chinese. Tests were conducted using high-quality recordings as nominal suspect samples and mobile-to-landline transmitted recordings as nominal offender samples. Performance was assessed before and after fusion with a baseline automatic mel frequency cepstral coefficient Gaussian mixture model universal background model (MFCC GMM-UBM) system. In addition, Monte Carlo simulations were performed to compare the output of the statistical models to true likelihood-ratio values calculated on the basis of the distribution specified for a simulated population. The formant-trajectory data of all speakers in the database was used to create a simulated population sample.

2. Methodology

2.1. Data

The data were extracted from a database of two non-contemporaneous voice recordings of each of 60 female speakers of Standard Chinese [22]. See [23] 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 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. The first and second recording sessions were separated by 2-3 weeks. High-quality recordings were made at 44 100 sampling frequency 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 [11, §2.2]. 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 were manually located and marked. There were between 6 and 41 stressed tokens of /iau/ per speaker per recording, median 21.5. /iau/ tokens were taken from realizations as a single word ("yao" one), 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).

2.2. Formant-trajectory measurement & parameterization

Human-supervised measurements of the trajectories of the first three formants (F1, F2, and F3) of each vowel token were made using FORMANTMEASURER [24]. See [12, 11] for details on the procedure for human-supervised formant-trajectory measurement. Discrete cosine transforms (DCTs) were fitted to the measured formant trajectories of all /iau/ tokens. See [8] for details of the procedure. In line with previous studies [11, 12, 21], the zeroth through fourth DCT coefficient values from F2 and F3 were used as variables in the present study.

2.3. Likelihood ratio calculation

2.3.1. Multivariate kernel density model

The multivariate kernel density (MVKD [6]) model assesses the difference between the samples taken from the suspect and the offender sample with respect to a background distribution estimated from a given population sample. While within-source variation is modeled by a Gaussian distribution, between-source variation is modeled using kernel density estimation. See [20] for a detailed discussion of the procedure with respect to its application in forensic voice comparison.

2.3.2. Principal component analysis kernel density LR model

In the principal component analysis kernel density likelihood ratio (PCAKLR [14]) model, problems with robustness of the MVKD model when using high-dimensional features are sidestepped by using PCA to obtain a decorrelating transform matrix and then computing the likelihood ratio as the product of univariate likelihood ratios of the projected features.

First, the mean of each feature is calculated from the suspect and offender samples and the samples of the speakers in the background sample and is then subtracted from each of the samples. The covariance $\mathbf{C}$ is then estimated from the mean-subtracted samples. The eigenvectors and eigenvalues of the covariance matrix are computed as in Eq. (1a). $\mathbf{V}$ is the matrix of eigenvectors $\mathbf{v}_m$ and $\mathbf{\Gamma}$ is a diagonal matrix composed of the eigenvalues γ_m corresponding to the eigenvectors $\mathbf{v}_m$ (off-diagonal values are zero). The suspect and offender samples and the samples of the speakers in the background sample are then transformed as in Eq. (1b).

$$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{\Gamma} \tag{1a}$$

$$\mathbf{y} = \mathbf{V}^{T}\mathbf{x} \tag{1b}$$

Univariate likelihood ratios are then computed for each dimension of the transformed features individually using a modified kernel density model given in Eq. (2) [15, p. 338],

$$\mathrm{LR} = \frac{K \exp\left(\frac{-(\bar{x}_s-\bar{x}_o)^2}{2a^2\sigma^2}\right) \sum_{i=1}^{k} \exp\left(\frac{-(m+n)(w-\bar{z}_i)}{2[\sigma^2+(m+n)s^2\lambda^2]}\right)}{\sum_{i=1}^{k} \exp\left(\frac{-m(\bar{x}_s-\bar{z}_i)}{2(\sigma^2+ms^2\lambda^2)}\right) \sum_{i=1}^{k} \exp\left(\frac{-n(\bar{x}_o-\bar{z}_i)}{2(\sigma^2+ns^2\lambda^2)}\right)} \tag{2a}$$

$$K = \frac{k\sqrt{m+n}\sqrt{\sigma^2+ms^2\lambda^2}\sqrt{\sigma^2+ns^2\lambda^2}}{a\sigma\sqrt{mn}\sqrt{\sigma^2+(m+n)s^2\lambda^2}} \tag{2b}$$

where $\bar{x}_s$ and $\bar{x}_o$ are the mean of the suspect and offender samples, $\bar{z}_i$ is the mean of speaker i in the background sample, k is the number of background speakers, σ^2 and s^2 are the within- and between-speaker variances, n and m are the number of tokens in the suspect and offender samples, $a^2 = \sqrt{1/m + 1/n}$ is the scaling factor of the within-source variance, $w = (n\bar{x}_s + m\bar{x}_o)/(m+n)$ is the weighted suspect and offender mean, and λ is the smoothing factor for the kernel density estimate.

The final likelihood ratio is then calculated as the product of the individual univariate likelihood ratios.

2.3.3. Multivariate normal model

The multivariate normal (MVN [6]) model assesses the difference between the samples taken from the suspect and the offender sample with respect to a background distribution estimated from a given population sample. Both within- and between-source variation are modeled by Gaussian distributions. See [15] for a detailed discussion of the procedure.

2.4. 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 [25]. Feature warping [26] using a three second sliding window was applied to the MFCCs and deltas before subsequent modeling. A GMM-UBM model [19] 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 256. Extraction of MFCCs and training of GMMs was performed using the Hidden Markov Toolkit [27].

2.5. 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 [28, 29, 30] which was applied to convert the scores from the test set to

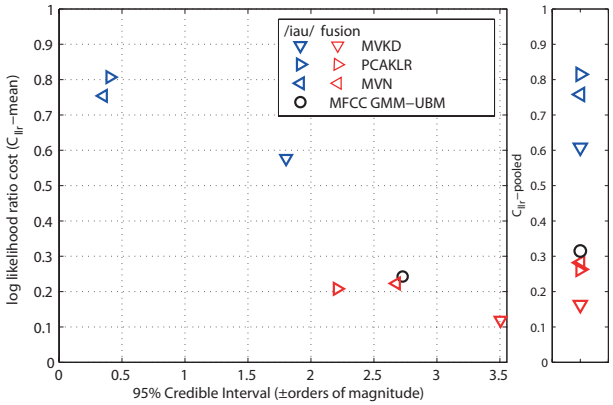

Figure 1: *Measures for validity (C_{llr}) and reliability (*$\log 10$ *95% credible interval) for systems using the MVKD, PCAKLR, and MVN models on /iau/ tokens individually (blue) as well as after fusion with the baseline MFCC GMM-UBM system (red).*

likelihood ratios. Logistic regression was also used to fuse the scores from multiple individual systems and convert them to likelihood ratios [31].

3. 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) [32] (C_{llr}-mean was calculated using the mean procedure and the 95%CI using the parametric procedure; C_{llr}-pooled gives the log likelihood ratio cost calculated using the pooled procedure). 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 [32, 20]. Results were also graphically represented using Tippett plots (for an introduction to the interpretation of Tippett plots see [32]).

Figure 1 shows the results for the mobile-to-landline v high-quality tests. The baseline MFCC GMM-UBM system had a C_{llr}-mean of 0.243 and a $\log_{10}$ 95% CI of 2.728. Fusion of the MVKD-based system with the baseline system resulted in a substantial improvement in validity (C_{llr}-mean 0.119, -51%) at a loss in reliability ($\log_{10}$ 95% CI 3.5, $+28\%$). Fusion of the PCAKLR-based system with the baseline system resulted in a smaller improvement in validity (C_{llr}-mean 0.208, -14%) at substantially improved reliability ($\log_{10}$ 95% CI 2.2, -19%). Fusion of the MVN-based system with the baseline system resulted in minor improvements in both validity (C_{llr}-mean 0.223, -8%) and reliability ($\log_{10}$ 95% CI 2.678, -2%). In terms of the pooled log likelihood ratio cost (C_{llr}-pooled) shown on the right in Figure 1, fusion of the MVKD-based system with the baseline system provided the highest improvement in performance, followed by the PCAKLR-based and the MVN-based systems.

Figure 2 shows Tippett plots of the baseline MFCC GMM-UBM system (left) and after fusion with systems based on the MVKD, PCAKLR, and MVN models. Reduction in C_{llr} appears to be primarily due to large magnitude log likelihood ratios supporting consistent-with-fact hypotheses getting even larger. For the MVKD-based system there is also a small decrease in 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.

The results reported here are of tests where the number of tokens per recording was not controlled. Restriction to six /iau/ tokens per recording session resulted in a large deterioration in performance (C_{llr} values before fusion of 0.9–1) and a similar, yet less pronounced, relative pattern of performance for the three statistical models.

4. Monte Carlo simulation

In practice, the true statistical distribution for a given population sample is not known. Following the approach in [34], Monte Carlo simulation is used to compare the output of the three statistical models to true likelihood-ratio values calculated on the basis of distributions specified for a simulated population.

In order not to diverge from the distribution of formant-trajectory-based measurements, a simulated population sample was generated based on the data samples of all speakers in the database. A set of 1000 simulated speakers were generated as follows: First, one recording session was randomly selected from the 120 recording sessions. Then, 10 tokens were uniformly sampled from the tokens from the high-quality recording (suspect condition) of that session. The parameters of a multivariate Gaussian are then estimated from that sample by calculating the mean and covariance of the sample. From this multivariate Gaussian, 30 observations were randomly generated as suspect-condition sample. Further, 10 tokens were uniformly sampled from the tokens from the mobile-to-landline transmitted recording (offender condition) of the other recording session of the same speaker, i.e., if the Session 1 recording was selected to generate the suspect-condition sample, then the Session 2 recording was selected to generate the offender-condition sample, and vice versa for Session 2 and Session 1. The parameters of a multivariate Gaussian are then estimated from that sample by calculating the mean of the sample and calculating the covariance as weighted average of the sample covariance and the pooled population covariance. From this multivariate Gaussian, 30 observations were randomly generated as offender-condition sample. Random numbers were generated using MATLAB's `mvnrnd` and `randi` functions. Of the 1000 simulated speakers, 100 were selected as test set and the remaining 900 as background set.

True likelihood ratios based on the distribution of the population specified for the simulation were calculated as follows:

$$LR = \frac{f(\bar{\mathbf{x}}_o|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)}{\frac{1}{J}\sum_{j=1}^{J} f(\bar{\mathbf{x}}_o|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{3}$$

where $\bar{\mathbf{x}}_o$ is the mean of the offender sample, $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ are the mean and covariance matrix computed from the randomly generated suspect sample, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the means and covariance matrices computed from each of $J = 900$ randomly generated suspect-condition speaker samples in the background set, and $f(\cdot)$ is the Gaussian probability density function. Similarly to the evaluation using real data, every test speaker's offender sample was compared with their own suspect-sample for a same-speaker comparison and with every other speaker's suspect sample as different-speaker comparisons. True likelihood ratios and likelihood-ratio estimates from the MVKD,

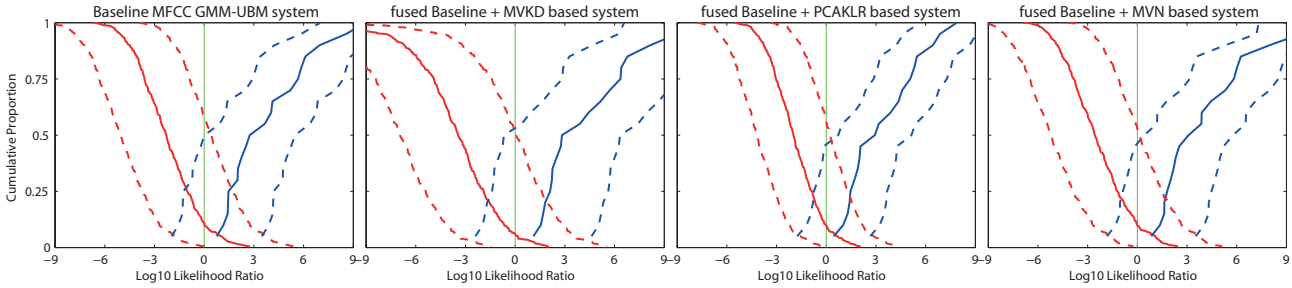

Figure 2: *Tippett plots of the baseline MFCC GMM-UBM system (left-most plot) and after fusion with systems based on the MVKD, PCAKLR, and MVN models (mobile-to-landline v high-quality recordings).*

Table 1: *Root-mean-square (RMS) deviation between estimated and true likelihood ratio values over simulated suspect and offender comparisons.*

		raw LR	calib. LR
MVKD	raw	47.54	81.57
	calib.	81.35	0.39
PCAKLR	raw	621.46	702.51
	calib.	81.63	1.14
MVN	raw	78.46	4.06
	calib.	81.18	0.54

PCAKLR, and MVN models were calculated for each speaker comparison.

The likelihood-ratio output of statistical models typically used in forensic voice comparison usually has to be calibrated (scores converted to likelihood ratios) before use [35, 29, 30]. For this evaluation, a set of *calibrated* likelihood ratios is computed in addition to the *raw* likelihood ratios by calibrating *raw* likelihood ratios using logistic-regression. Calibration weights were estimated using leave-one-out crossvalidation from the raw likelihood ratios.

Table 1 shows the root-mean-square (RMS) deviation between true log-likelihood-ratio values (*raw* and *calibrated*) and estimated log-likelihood-ratio values (*raw* and *calibrated*) of simulated suspect and offender comparisons. Raw likelihood-ratio values show a large deviation, particularly for PCAKLR, highlighting the importance of calibration. After calibration, the RMS deviation for all three models is very similar (about 81). When comparing the calibrated likelihood-ratio outputs with the calibrated true likelihood ratio values, the MVKD-based system shows the lowest deviation, closely followed by the MVN and PCAKLR-based models.

5. Discussion & Conclusions

The present paper assessed the performance of multivariate kernel density (MVKD), the principal component analysis kernel density likelihood ratio (PCAKLR), and multivariate normal (MVN) models for likelihood-ratio calculation in acoustic-phonetic forensic-voice-comparison systems. Each method was applied to the same set of data. The data were coefficient values obtained from discrete cosine transforms fitted to human-supervised formant-trajectory measurements of tokens of /iau/ from a database of recordings of 60 female speakers of Chinese. Tests were conducted using high-quality recordings as nominal suspect samples and mobile-to-landline transmitted recordings as nominal offender samples. Performance was assessed as degree of improvement over a baseline MFCC GMM-UBM system. Overall, the MVKD model provided the highest improvement in validity, while reliability deteriorated. PCAKLR showed small improvements in validity and sizable improvement in reliability. The multivariate normal (MVN) model showed only minor improvements in both validity and reliability.

Monte Carlo simulations were used to compare the output of the statistical models to true likelihood-ratio values calculated on the basis of the distribution specified for a simulated population. When raw likelihood-ratio scores were calibrated, none of the methods clearly outperforms the others; however, the MVKD model showed the lowest RMS deviation when both its output and the true likelihood-ratio values were calibrated.

The present study only tested one phonetic unit (/iau/) in recordings of female speakers using one speaking style under a specific mismatch condition. 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.

6. References

[1] I. W. Evett, C. G. G. Aitken, C. E. H. Berger, J. S. Buckleton, C. Champod, J. M. Curran, A. P. Dawid, P. Gill, J. González-Rodríguez, G. Jackson, A. Kloosterman, T. Lovelock, D. Lucy, P. Margot, L. McKenna, D. Meuwly, C. Neumann, N. NicDaéid, A. Nordgaard, R. Puch-Solis, B. Rasmusson, M. Redmayne, P. Roberts, B. Robertson, C. Roux, M. J. Sjerps, F. Taroni, T. Tjin-A-Tsoi, G. A. Vignaux, S. M. Willis, and G. Zadora, "Expressing evaluative opinions: a position statement," *Sci. Justice*, vol. 51, pp. 1–2, 2011.

[2] C. E. H. Berger, J. Buckleton, C. Champod, I. W. Evett, and G. Jackson, "Evidence evaluation: A response to the Court of Appeal judgment in R v T," *Sci. Justice*, vol. 51, pp. 43–49, 2011.

[3] B. Robertson, G. A. Vignaux, and C. E. H. Berger, "Extending the confusion about Bayes," *Mod. L. Rev.*, vol. 74, pp. 444–455, 2011.

[4] B. Robertson and G. A. Vignaux, *Interpreting evidence*. Chichester, UK: Wiley, 1995.

[5] C. C. G. Aitken, P. Roberts, and G. Jackson, "Fundamentals of probability and statistical evidence in criminal proceedings: guidance for judges, lawyers, forensic scientists and expert witnesses," in *Practitioners Guide No 1, Royal Statistical Society's Working Group on Statistics and the Law*, 2010.

[6] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics*, vol. 53, no. 1, pp. 109–122, 2004.

[7] Y. Kinoshita, S. Ishihara, and P. Rose, "Exploring the discriminatory potential of f0 distribution parameters in traditional forensic speaker recognition," *Int. J. of Speech, Lang. and the Law*, vol. 16.1, pp. 91–112, 2009.

[8] G. S. Morrison, "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2387–2397, 2009.

[9] P. Rose, "The effect of correlation on strength of evidence estimates in forensic voice comparison: uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics," *Int. J. Biometrics*, vol. 2, no. 4, pp. 316–329, 2010.

[10] ——, "More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends," *Int. J. Speech Lang. Law*, vol. 20, no. 1, pp. 77–116, 2013.

[11] 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.

[12] 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.

[13] G. S. Morrison and E. Enzinger, "Forensic speech science – Review: 2010–2013," in *Proceedings of the 17th International Forensic Science Managers' Symposium*, N. NicDaéid, Ed. Lyon, France: INTERPOL, 2013, pp. 616–623, 629–635.

[14] B. Nair, E. Alzqhoul, and B. J. Guillemin, "Determination of likelihood ratios for forensic voice comparison using principal component analysis," *Int. J. Speech Lang Law*, vol. 21, no. 1, pp. 83–112, 2014.

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

[16] P. Rose, D. Lucy, and T. Osanai, "Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects model - a non-idiot's bayes' approach," in *Proceedings of the 10th Australasian International Conference on Speech Science and Technology (SST-2004)*, 2004.

[17] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech and Language*, vol. 20, pp. 159–191, 2006.

[18] E. Enzinger and C. H. Kasess, "Bayesian vocal tract model estimates of nasal stops for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 1685–1689.

[19] 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.

[20] 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.

[21] C. Zhang, G. S. Morrison, and T. Thiruvaran, "Forensic voice comparison using Chinese /iau/," in *Proc. ICPhS XVII*, Hong Kong, China, 2011, pp. 2280–2283.

[22] 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/

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

[24] G. S. Morrison and T. M. Nearey. (2011) FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories. [Online]. Available: http://geoff-morrison.net/#FrmMes

[25] 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.

[26] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey 2001: The Speaker Recognition Workshop*, Crete, Greece, 2001, pp. 213–218.

[27] S. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Department of Engineering, Cambridge University, U.K., Tech. Rep., 1993.

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

[29] 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.

[30] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Aus. J. Forensic Sci.*, vol. 45, pp. 173–197, 2013.

[31] 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.

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

[33] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1895–1898.

[34] G. S. Morrison, "Calculation of forensic likelihood ratios: Use of Monte Carlo simulations to compare the output of score-based approaches with true likelihood-ratio values," submitted.

[35] D. Ramos-Castro, J. González-Rodríguez, and J. Ortega-Garcia, "Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework," in *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.