Accepted Manuscript

Title: The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings

Author: Geoffrey Stewart Morrison

PII: S0379-0738(17)30540-6
DOI: https://doi.org/10.1016/j.forsciint.2017.12.024
Reference: FSI 9106

To appear in: *FSI*

Received date: 30-10-2017
Revised date: 11-12-2017
Accepted date: 12-12-2017

Please cite this article as: Geoffrey Stewart Morrison, The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings, Forensic Science International https://doi.org/10.1016/j.forsciint.2017.12.024

Version 2017-12-11a

The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings

Geoffrey Stewart Morrison [a,b,*]

[a] Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom

[b] Forensic Evaluation Ltd, Birmingham, England, United Kingdom

[*] Corresponding author. *E-mail address*: geoff-morrison@forensic-evaluation.net

Version 2017-12-11a

Highlights

- The performance of a GMM-UBM forensic voice comparison system was tested.
- Using high-quality audio to train the UBM, and without calibration.
- Using known-speaker-condition audio to train the UBM, and with calibration.
- The first variant was that used by a practitioner in a case.
- Its performance was very poor.

Abstract

In a 2017 New South Wales case, a forensic practitioner conducted a forensic voice comparison using a Gaussian mixture model - universal background model (GMM-UBM). The practitioner did not report the results of empirical tests of the performance of this system under conditions reflecting those of the case under investigation. The practitioner trained the model for the numerator of the likelihood ratio using the known-speaker recording, but trained the model for the denominator of the likelihood ratio (the UBM) using high-quality audio recordings, not recordings which reflected the conditions of the known-speaker recording. There was therefore a difference in the mismatch between the numerator model and the questioned-speaker recording versus the mismatch between the denominator model and the questioned-speaker recording. In addition, the practitioner did not calibrate the output of the system. The present paper empirically tests the performance of a replication of the practitioner's system. It also tests a system in which the UBM was trained on known-speaker-condition data and which was empirically

calibrated. The performance of the former system was very poor, and the performance of the latter was substantially better.

Keywords: Forensic voice comparison; Automatic speaker recognition; GMM-UBM; Likelihood ratio; Validation; Calibration; Admissibility

1 Introduction

The present paper reports on an empirical evaluation of the performance of different variants of a forensic voice comparison system. One variant is that used by a forensic practitioner in a case, and another variant is that recommended by another forensic practitioner who critiqued the first practitioner's report. The second practitioner was the author of the present paper.

The case was a 2017 case in New South Wales (NSW), Australia. Besides the technical issues raised in the present paper, the case is of interest because it involved an attempt to have testimony based on an automatic approach to forensic voice comparison admitted. To my knowledge, testimony based on this approach has not yet been admitted in an Australian jurisdiction. An attempt to have testimony based in-part on an automatic approach was made in US Federal Court in 2015 [1][2]. In that instance, a *Daubert* hearing was held, but, before the judge issued a ruling on admissibility, the case was settled by plea deal. An attempt to have testimony based in-part on an automatic approach was also made in England & Wales in 2015 [3][4]. In that instance, the court was not satisfied that adequate and sufficient data had been used in training the system or in empirically testing its performance. Concerns were raised regarding the amount of data and whether the data reflected the relevant population. The testimony was ruled inadmissible as new evidence in the context of an appeal.

The practitioner in the 2017 NSW case did not provide results of an empirical evaluation of the performance of the forensic voice comparison system under conditions reflecting those of the case under investigation. Unlike US Federal Rule of Evidence (FRE) 702 and the *Daubert* criteria [5][6], or England & Wales Criminal Practice Directions (CPD) 19A [7], the NSW version of the Australian Uniform Evidence Act [8] does not require demonstration of "scientific validity" (*Daubert* fn 9). From a scientific perspective, however, empirical validation under conditions sufficiently similar to those of the case is the only way to know how well the system works under those conditions. [9] reviewed calls going back to the 1960s for the performance of forensic voice comparison systems to be empirically tested under casework conditions, and, with respect to feature-comparison methods in general, the point was recently reiterated by President Obama's Council of Advisors on Science and Technology [10][11].

In the 2017 NSW case, I submitted an initial critique of the first practitioner's report. The critique identified some theoretical problems with the way the forensic voice comparison software had been used. Whether those theoretical problems would lead to substantial practical problems, was, however, unclear. I replicated the practitioner's forensic voice comparison system in order to empirically test its

performance. The plan was to compare the performance of the variant of the system used by the practitioner with a variant in which the theoretical problems had been fixed. This would either reveal that the theoretical problems had negligible practical relevance, or that they had a substantial impact on actual system performance. If the former were true, then these particular theoretical criticisms would be moot. If the latter were true, then the practical demonstration of the effect would be potentially much more convincing to the court than a theoretical argument alone.

After I submitted my initial critique, and after I replicated the system, but before I conducted the empirical tests of the performance of the system, and before the case went to trial, the case was resolved via a plea deal. I expect, however, that there will be an increasing number of attempts to have automatic approaches to forensic voice comparison admitted in various jurisdictions, and that the discussion I present in the present paper will therefore be of broader interest. Below, I describe the relevant conditions of the case, discuss the theoretical problems, describe the system, and present the results of empirical tests of variants of the replicated system.

2 Description of the case

The conditions of the questioned-and known-speaker recordings were mismatched, and both were poor.

The questioned-speaker recording was of a mobile telephone call made to emergency services. The telephone was wedged under a mattress which was resting directly on a carpeted floor (the microphone may have been in a part of the telephone protruding from under the mattress). The speaker of interest was distant from the microphone.

The known-speaker recording was of a landline call made from a jail. The call was made in a highly reverberant environment, with background noise including other speakers' loud voices, and it was saved in a lossy compressed format (MP3).

In each recording, the speaker of interest was a male who spoke English with an Australian accent.

3 Description of the analysis conducted by the forensic practitioner

The forensic practitioner performed a forensic voice comparison using a Gaussian mixture model - universal background model (GMM-UBM) system implemented using the open-source Microsoft Research (MSR) Identity Toolkit version 1.0 [12][13]. The Voicebox Toolbox [14] was used for feature extraction.

The GMM-UBM modelling approach is a standard approach which has been in use for almost 20 years. It was used in the previous generation of commercially marketed forensic voice comparison systems, but has been mostly supplanted by a newer modelling approach (called "i-vector PLDA"), and state of the art in automatic speaker recognition research has moved on to an even newer modelling approach (based on "deep neural networks", DNNs). As mature technology, however, GMM-UBM is unobjectionable. The issue of relevance for the court should not be whether it is the best performing approach, but whether its performance is good enough under the conditions of the case. The features used as input to the GMM-

UBM system in this case (MFCCs + deltas + double deltas, see below) are also standard, mature, and unobjectionable.[1]

In essence, in order to calculate the numerator of a likelihood ratio, a probability density model is trained on the feature values extracted from the known-speaker recording, and in order to calculate the denominator of the likelihood ratio, a probability density model is trained on the feature values extracted from recordings of a sample of speakers representative of the relevant population specified in the defence hypothesis (the latter model is the UBM). The likelihood of each model is then assessed at the feature values extracted from the questioned-speaker recording, and the likelihood of the numerator is divided by the likelihood of the denominator. The likelihood ratio is intended to answer the question: What is the probability of obtaining the measured acoustic properties of the voice on the questioned-speaker recording if it were produced by the known speaker, versus what is the probability of obtaining the measured acoustic properties of the voice on the questioned-speaker recording if it were produced not by the known speaker but by some other speaker selected at random from the relevant population?

The remainder of this section provides technical details assuming familiarity with the GMM-UBM approach. The details are based on information in the practitioner's report, responses from the practitioner to a request for technical details sufficient to allow for replication of the system, and publically available information about the data and software used by the practitioner.

Manual diarization plus a log-energy-based voice activity detector (VAD) were used to isolate the speech of the speaker of interest on each recording. A total of 72 s of speech was extracted from the known-

[1] The reviewers of the present paper actually did object to the use of the GMM-UBM system on the grounds that it could not be expected to provide sufficient mismatch compensation and that a system such as an i-vector PLDA system would include better mismatch compensation and would be expected to produce better results. In a study reported in [15] (published after I worked on the case in question and after the first draft of the present paper was prepared), an i-vector PLDA system and a GMM-UBM system were compared – both systems were implemented using the MSR Toolkit, and both were trained and tested on the *foresnic_eval_01* data. In that study it was found that the i-vector PLDA system did in fact outperform the GMM-UBM system. That there exists some other system that is expected to perform better, or that another system is even demonstrated to perform better, cannot, however, be used as an argument against legal admission. Legal admissibility standards such as FRE 702 - *Daubert* or CPD 19A require the demonstration of the level performance of the system actually used in the case (under conditions reflecting those of the case), and consideration of whether that demonstrated level of performance is sufficient. Neither FRE 702 - *Daubert* nor CPD 19A mentions consideration of whether some other system may outperform the one actually used in the case. Given the pace of progress in research, there will always be some other system that could be expected to perform better. It could be argued that a DNN system would outperform an i-vector PLDA system, or that one DNN system would outperform another. If such arguments were relevant then it is unlikely that any system in any branch of forensic science would be admitted in any case. It may be that an i-vector PLDA system outperforms a GMM-UBM system, and that if presented with both (e.g., one proffered by the prosecution and one by the defence) a court could decide that the former has demonstrated a sufficient level of performance to warrant admission but that the latter has not, or if both are admissible that the output of the former be given more weight. But the decision as to admissibility should be made on an absolute performance criterion, not a relative one; and if the court is only presented with a single system, it need only be concerned with whether the demonstrated level of performance of that system satisfies the criterion.

speaker recording, and 63 separate utterances totalling 280 s of speech were extracted from the questioned-speaker recording. Log energy + 12 mel frequency cepstral coefficients (MFCCs) + deltas + double deltas [16][17] were extracted once every 10 ms using a 20 ms wide window. This resulted in feature vectors with 39 dimensions. Cepstral mean and variance normalization (CMNV) was applied as a feature-domain mismatch compensation technique [18][19]. CMNV for each recording was trained on all feature vectors of that recording that remaining after VAD.

Recordings of 71 male speakers from the AusTalk database [20] were used to train the UBM. These were high-quality audio recordings, recorded using a head mounted microphone while the speakers were engaged in the "map task" (a task in which speakers have to ask and give directions while looking at a map). The speakers had Australian English accents. The UBM had 1024 Gaussian components with diagonal-only covariance matrices. The known-speaker model was adapted from the UBM using maximum a posteriori (MAP) adaptation [12].

The number of speakers on the questioned-speaker recording was disputed, so a likelihood ratio was calculated separately for each utterance. Given 63 utterances totalling 280 s, the mean length of an utterance would have been ~4.44 s. The likelihood of the known-speaker model and the likelihood of the UBM were evaluated at each feature vector from the questioned-speaker recording, the former likelihood was divided by the latter, and the mean of the resulting log likelihood ratios was calculated over all feature vectors in an utterance. The exponent of the mean of the log likelihood ratios was then reported as a "likelihood ratio".

The resulting per-utterance "likelihood ratio" values ranged from 1.6 to 5.5. The 10th, 50th, and 90th percentiles were 1.9, 2.4, and 3.7 respectively.

4 Theoretical problems

My initial critique raised a number of concerns. The concerns which will be empirically addressed in the present paper are the following.

First, the AusTalk recordings used to train the model for the denominator of the likelihood ratio (the UBM), were high-quality audio recordings, whereas the known-speaker recording, used to train the model for the numerator of the likelihood ratio, had background noise and reverberation, was transmitted via a landline telephone system, and was saved in a lossy compressed format. The likelihoods of these models were assessed at feature values extracted from the questioned-speaker recording, a distant-speaker mobile-telephone-transmitted recording. The mismatch between the questioned-speaker data and the data used to train the numerator model was therefore different from the mismatch between the questioned-speaker data and the data used to train the denominator model. I expected this to introduce a bias: either the numerator or the denominator would be larger because the conditions of the data used to train the model for one or the other will be closer to the conditions of the questioned-speaker data. To avoid this bias, my recommendation was to train the UBM using recordings which reflect the conditions of the known-speaker recording [21]. That way the mismatch between the numerator model and the questioned-speaker data, and the mismatch between the denominator model and the questioned-speaker

data would be the same. CMVN is intended to compensate for mismatched conditions, but, given the extreme differences between the conditions of the known-speaker recording and the recordings from the AusTalk database, I was sceptical that this technique would be successful in sufficiently ameliorating the potential bias. These theoretical concerns could be empirically investigated by training two GMM-UBM models, one using high-quality data to train the UBM and the other using data with conditions similar to those of the known-speaker data.

Second, the GMM-UBM implemented by the practitioner required estimation of 80,895 parameter values (39 feature dimensions × (1024 mean vectors + 1024 variances) + 1023 weights). Obtaining good estimates for such a large number of parameters would require an enormous amount of training data. In addition, MFCCs values came from a series of adjacent frames, and were not therefore statistically independent. Hence, naïve Bayes fusion, i.e., multiplying likelihood ratio values (or summing log likelihood ratio values) from adjacent frames, would not produce an appropriate likelihood ratio value for the combination of the individual likelihood ratios. An additional complication is that the values reported were not the exponent of the sum of the individual log likelihood ratios but the exponent of the mean of the individual log likelihood ratios. These would not be appropriate values for a combined likelihood ratio even if the individual likelihood ratios were statistically independent. The absolute value output by such a model is therefore generally not considered directly interpretable as a meaningful answer to the question specified by the prosecution and defence hypotheses. The standard solution to this problem is to not treat the output of GMM-UBM models as interpretable likelihood ratio values, but as *scores* which must be converted to likelihood ratios (or *calibrated*) before they can be interpreted [22]–[25]. Whether in practice this would result in a substantial difference in the reported strength of evidence could be empirically investigated by adding a calibration stage to the system and comparing the raw score output of the GMM-UBM with calibrated likelihood ratios derived from those scores.

5 Empirical testing

I did not have access to a database of recordings reflecting the particular conditions of the known- and questioned-speaker recordings in this case (and prior to the case I was not provided with the resources or time necessary to obtain such data). The empirical tests reported in the present paper instead use data from a multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) [26]. These data have a questioned-speaker condition which is a landline telephone call with background office noise (babble and typing noises) and saved in a lossy compressed format, and a known-speaker condition which is a direct microphone recording but with reverberation and ventilation system noise. The conditions differ from those of the case under consideration, and the results of tests using these data are therefore not directly informative of the expected performance of the GMM-UBM system under those conditions. If, however, when using the *forensic_eval_01* data, a substantial difference in performance is found between the variant of the system used by the practitioner and the variant that includes the changes I recommended, this would increase concern about the performance of that first variant under the conditions of the case. I would argue that it was incumbent upon a practitioner to demonstrate the level of performance of their system under

conditions reflecting those of the case under consideration. I would argue this in general, and also if the practitioner in this case had wanted to claim that any problems empirically demonstrated using the *forensic_eval_01* data would not occur under the conditions of the case.

I trained the UBM using one known-speaker-condition recording from each of the 105 speakers in the *forensic_eval_01* training set. The number of feature vectors extracted from each recording ranged from 7,575 to 11,497 with a median of 10,248 (equivalent of 76 to 115 s of speech with a median of 102 s). The known-speaker conditions in the *forensic_eval_01* data had been created by applying signal processing techniques to high-quality recordings from a database of recordings of 500+ Australian English speakers [27][28]. See [21] for details of the signal-processing procedures that were applied. From the database of recordings of 500+ Australian English speakers I obtained the original high-quality versions of the training-set recordings, and trained another UBM using the parallel set of feature vectors extracted from those recordings.

The *forensic_eval_01* test set included one recording of each of 61 speakers in the questioned-speaker condition, plus two or three known-speaker-condition recordings from each speaker. The questioned-speaker-condition recording came from the first of up to three non-contemporaneous recording sessions, and each of the known-speaker-condition recordings came from a different session. The number of feature vectors extracted from each known-speaker-condition recording ranged from 9,367 to 10,738 with a median of 10,125 (equivalent of 94 to 107 s of speech with a median of 101 s). Using each known-speaker-condition recording in the test set, an individual-speaker model was adapted from the UBM. A score was then calculated for the combination of each questioned-speaker-condition recording and each individual-speaker model, with the exception that for same-speaker comparisons the questioned-speaker-condition recording was not compared with the contemporaneous first-session individual-speaker (known-speaker-condition) model. Rather than base calculations on all of the feature vectors from each questioned-speaker-condition recording in the test set, I used contiguous blocks of 444 feature vectors, equivalent to 4.44 s of speech, i.e., the mean duration of the questioned-speaker utterances in the case under consideration. The number of utterance-length blocks ranged from 5 to 9 per questioned-speaker-condition recording, with a median of 7.

A cross-validated procedure was adopted for training a calibration model (score to likelihood ratio conversion model). When a same-speaker score was to be calibrated, all scores which had been calculated using any recording of that speaker as a member of the known- versus questioned-condition pair were held out from model training, and when a different-speaker score was to be calibrated, all scores which had been calculated using any recording of either of those two speakers as a member of the known- versus questioned-condition pair were held out from model training. The calibration model used was logistic regression. This is a commonly used model for score to likelihood ratio conversion [22]–[25].

The results of the empirical tests of the performance of different variants of the GMM-UBM system are provided graphically as Tippett plots in Figures 1 through 4, and numerically as log likelihood ratio costs (C_{llr}) in Table 1. For explanations of these graphics and metrics, see [24],[26],[29]–[33].

The performance of the variant of the system that used high-quality audio data to train the UBM and did not include calibration (the variant of the system used by the practitioner in the case) was very poor. The better the performance of a system, the smaller the value of the C_{llr}. A system which gives no useful information, a system that always outputs a likelihood ratio of 1 irrespective of the input, will have a C_{llr} of 1. The first variant of the system had a C_{llr} very close to 1, indicating that it gave almost no useful information. Almost all of the "likelihood ratios" output by this variant had values above 1 (albeit just above 1), irrespective of whether they were from same-speaker or different-speaker comparisons (see Fig. 1). As I predicted on the basis of theory, this variant has a bias.

Turning to the system variant that used known-speaker-condition data to train the UBM but did not include calibration. That system did not have the same bias as in the first system (compare Fig. 2 with Fig. 1). Comparing the parallel "likelihood ratios" output by the two variants, the values output by the first variant were on average (mean) 18% higher than those output by the second variant. The performance of the second variant was, however, still very poor: C_{llr} was close to 1 and all the "likelihood ratios" output by this variant had values close to 1. That the second variant offers a potential improvement over the first system can, however, be discerned by considering C_{llr}^{min}. This is the same as C_{llr}, but calculated after a non-parametric calibration process (pool adjacent violators) has been applied to the output of each system variant. The procedure involves training and testing on the same data, so it cannot be used to quantify expected performance on new data, but it does allow us to determine whether there is any difference in discriminating power between the two system variants with respect to the data on which they were both tested (they were both tested on exactly the same data). In Table 1 we see that C_{llr}^{min} is substantially lower for the second variant than for the first variant. Careful inspection of Fig. 1 and Fig. 2 also reveals that, in addition to a lack of bias in the second variant, the cross-over point between the same-speaker and different-speaker curves (the equal error rate) is lower for the second variant.

Using high-quality audio data to train the UBM, but then calibrating using data that reflect the conditions of the known- and questioned-speaker recordings resulted in better performance than either of the first two variants. C_{llr} is somewhat lower, the Tippet plot in Fig. 3 does not reveal a bias, and the likelihood ratio values output by this variant of the system can reach higher and lower values.

Using audio data reflecting the conditions of the known-speaker recording to train the UBM, and calibrating using data that reflect the conditions of the known- and questioned-speaker recordings, the variant of the system I recommended on theoretical grounds, had much better performance than any of the other variants.[2] C_{llr} is substantially lower, the Tippett plot in Fig. 3 does not reveal a bias, and the

[2] The performance is not particularly good and it could be decided that the level of performance of this system would not be sufficient to warrant admission in a case for which the conditions of the test data used in the present paper reflected those of the case. Poor performance is not unexpected given the conditions, which included the questioned-speaker-condition recordings having only 4.44 s worth of data (the results in [15] differ from those reported here because the latter used longer questioned-speaker-condition recordings and different feature-extraction options). The conditions of the 2017 NSW case were different from those of the test data used in the present paper, hence these results do not indicate what the performance of this

likelihood ratio values output by this variant can reach much higher and much lower values.[3]

6 Discussion and conclusion

The results of testing different variants of a GMM-UBM system illustrate the importance of training the model for the denominator of the likelihood ratio using data which reflect the conditions of the known-speaker recording, so that there is not a different mismatch between the denominator model and the questioned-speaker recording than between the numerator model and the questioned-speaker recording. The results also illustrate the importance of calibrating the output of a forensic voice comparison system using data that reflect the conditions of the known- and questioned-speaker recordings. Training using data reflecting the conditions of the case and calibrating system output should be standard practice.

The theoretical arguments and empirical test results presented in the present paper suggest that the performance of the GMM-UBM system variant employed by the practitioner in the 2017 NSW case would have been poor, but they do not test what performance under the particular conditions of the case would actually have been. I would argue, however, that these theoretical arguments and empirical test results should raise substantial doubt about whether the performance of the system would have been acceptable when applied to the known- and questioned-speaker recordings in the case. I would also argue that it is incumbent on the proposer of testimony to empirically test the level of performance of their forensic analysis system under conditions that are sufficiently similar to the conditions of the case under consideration that the results of those tests would be meaningful for deciding whether the performance of the system is acceptable for use in that case.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All opinions expressed in this document are those of the author and do not necessarily reflect the opinions or policies of any other individuals or organisations with which the author has been or is currently associated.

system would be under the conditions of the 2017 NSW case.

[3] One could be tempted to look for "likelihood ratio" values output by the first variant that are close to values reported by the practitioner for the analysis in the case, then find the values of the parallel likelihood ratios output by the third variant. This, however, would be misleading because the data used here to train the UBM and the calibration model are substantially different from the data used by the practitioner and substantially different from the conditions of the case. The feature-vector to score and the score to likelihood-ratio functions derived here are therefore not the same as would be derived using the data that the practitioner used or using data reflecting the conditions of the case.

References

[1] *United States v Ahmed et al*, No. 12-661 (E.D.N.Y.)

[2] G.S. Morrison, W.C. Thompson, Assessing the admissibility of a new generation of forensic voice comparison testimony, Columbia Sci. Tech. Law Rev., 18 (2017) 326–434.

[3] *R v Slade* [2015] EWCA Crim 71.

[4] G.S. Morrison, Admissibility of forensic voice comparison testimony in England and Wales, Crim. Law Review, (2018) 20–33.

[5] United States Federal Rule of Evidence 702 (as amended Apr. 17, 2000, eff. Dec. 1, 2000; Apr. 26, 2011, eff. Dec. 1, 2011).

[6] *Daubert et al v Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993).

[7] *Criminal Practice Directions* [2015] EWCA Crim 1567 (Consolidated with Amendments No. 2 [2016] EWCA Crim 1714, No. 3 [2017] EWCA Crim 30, No. 4 [2017] EWCA Crim 310, No. 5 [2017] EWCA Crim 1076).

[8] Evidence Act 1995 (*NSW*).

[9] G.S. Morrison, Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison, Sci. Just. 54 (2014) 245–256, http://dx.doi.org/10.1016/j.scijus.2013.07.004.

[10] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, 2016, https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/.

[11] E.S. Lander, Response to the ANZFSS council statement on the President's Council of Advisors on science and technology report, Aus. J. Forensic. Sci. 49 (2017) 366–368, http://dx.doi.org/10.1080/00450618.2017.1304992.

[12] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Process., 10 (2000) 19–41, http://dox.doi.org/10.1006/dspr.1999.0361.

[13] S.O. Sadjadi, M. Slaney, L. Heck, MSR Identity Toolbox v1.0: A MATLAB Toolbox for speaker recognition research, IEEE Speech Lang. Process. Tech. Committee Newsletter, November 2013, http://archive.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-11/IdentityToolbox/.

[14] M. Brookes, Voicebox: Speech processing toolbox for Matlab, 1997, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[15] G.D. da Silva, C.A. Medina, Evaluation of MSR Identity Toolbox under conditions reflecting those of a real forensic case (*forensic_eval_01*), Speech Commun. 94 (2017) 42–49, http://dx.doi.org/10.1016/j.specom.2017.09.001.

[16] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process., 28 (1980) 357–366, http://dox.doi.org/10.1109/TASSP.1980.1163420.

[17] S. Furui, Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust. Speech Signal Process. 34 (1986) 52–59, http://dox.doi.org/10.1109/TASSP.1986.1164788.

[18] S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29 (1981) 254–272, http://dox.doi.org/10.1109/TASSP.1981.1163530.

[19] O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Commun. 25 (1998) 133–147, http://dox.doi.org/10.1016/S0167-6393(98)00033-8.

[20] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. Lewis, A. Butcher, J. Hajek (2011). Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box, Proc. Interspeech, 2011, pp. 841–844.

[21] E. Enzinger, G.S. Morrison, F. Ochoa, A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Sci. Just. 56 (2016) 42–57, http://dx.doi.org/10.1016/j.scijus.2015.06.005.

[22] S. Pigeon, P. Druyts, P. Verlinde, Applying logistic regression to the fusion of the NIST'99 1-speaker submissions, Digital Signal Process. 10 (2000) 237–248, http://dx.doi.org/10.1006/dspr.1999.0358.

[23] D. Ramos Castro, Forensic evaluation of the evidence using automatic speaker recognition systems. Doctoral dissertation, Autonomous University of Madrid, 2007.

[24] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2104–2115, http://dox.doi.org/10.1109/TASL.2007.902747.

[25] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio, Austral. J. Forensic Sci. 45 (2013) 173–197, http://dx.doi.org/10.1080/00450618.2012.733025.

[26] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) – Introduction, Speech Commun. 85 (2016) 119–126, http://dx.doi.org/10.1016/j.specom.2016.07.006.

[27] G.S. Morrison, P. Rose, C. Zhang, Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, Austral. J. Forensic Sci. 44 (2012) 155–167,

http://dx.doi.org/10.1080/00450618.2011.630412.

[28] G.S. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B.K. Folkes, S. De Souza, N. Cummins, D. Chow, D., Forensic database of voice recordings of 500+ Australian English speakers, 2015, http://databases.forensic-voice-comparison.net/.

[29] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, Doctoral dissertation, University of Lausanne, 2001.

[30] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang. 20 (2006) 230–275, http://dox.doi.org/10.1016/j.csl.2005.08.001.

[31] G.S. Morrison, E. Enzinger, C. Zhang, Forensic speech science, in: I. Freckelton, H. Selby (Eds.), Expert Evidence, Thomson Reuters Sydney, Australia, 2017, ch. 99.

[32] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Just. 51 (2011) 91–98, http://dx.doi.org/10.1016/j.scijus.2011.03.002.

[33] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Internat. 2016, http://dx.doi.org/10.1016/j.forsciint.2016.03.048.

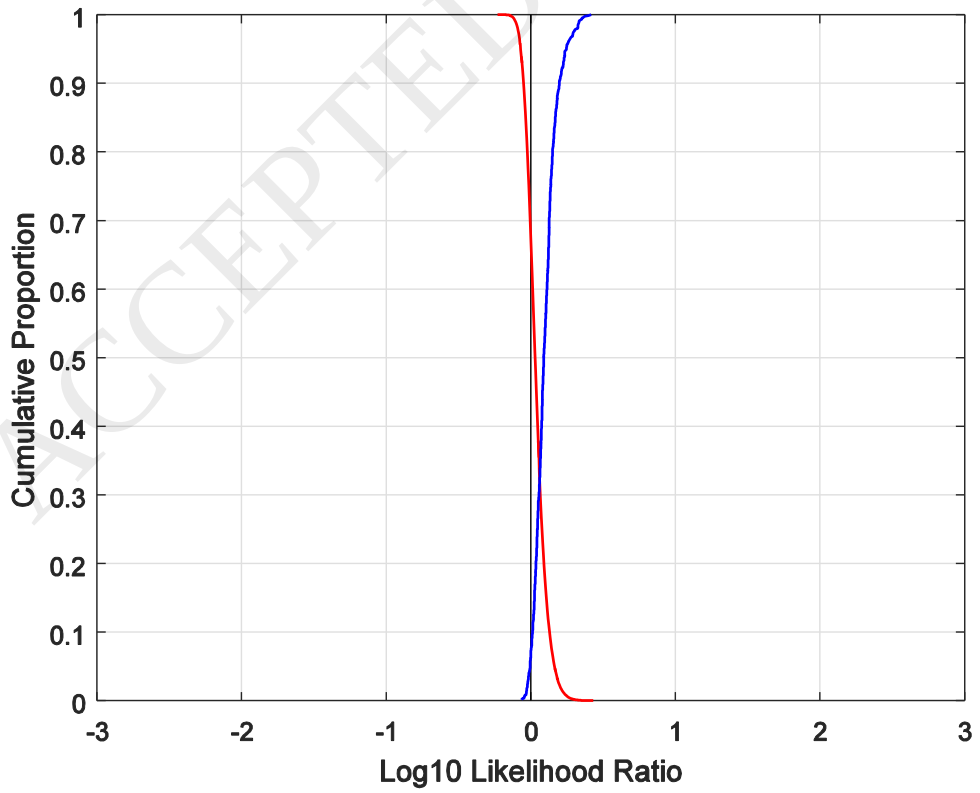
7 Figure captions

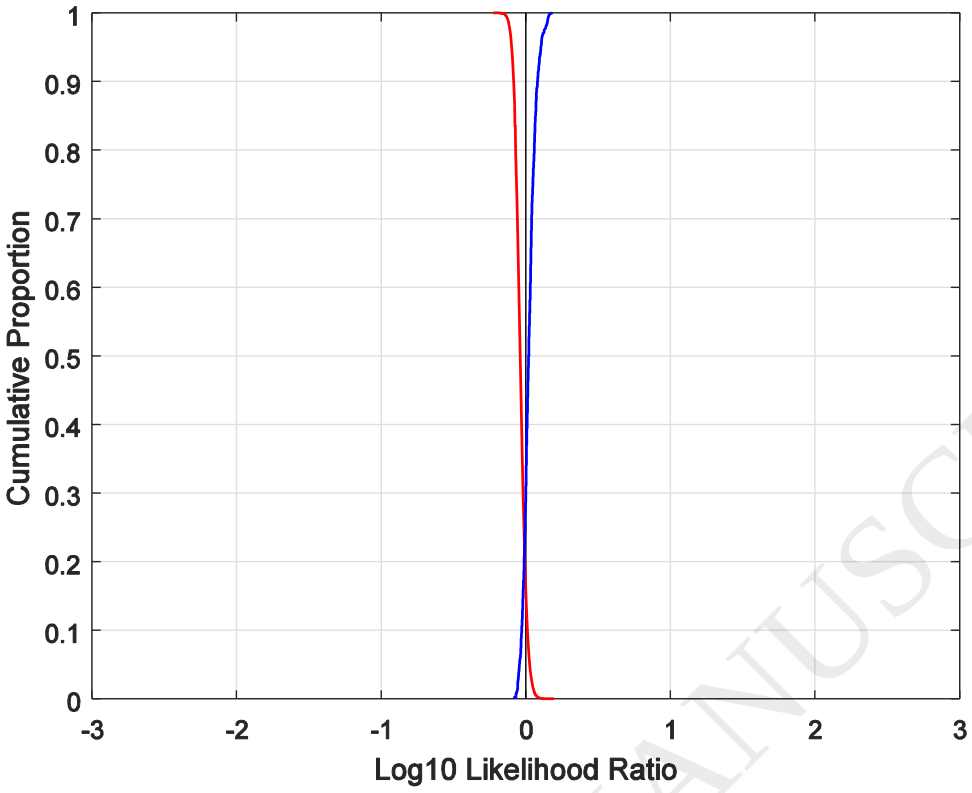
Fig. 1. Tippett plot of the results of testing the system variant with high-quality audio data used to train the UBM and no calibration.

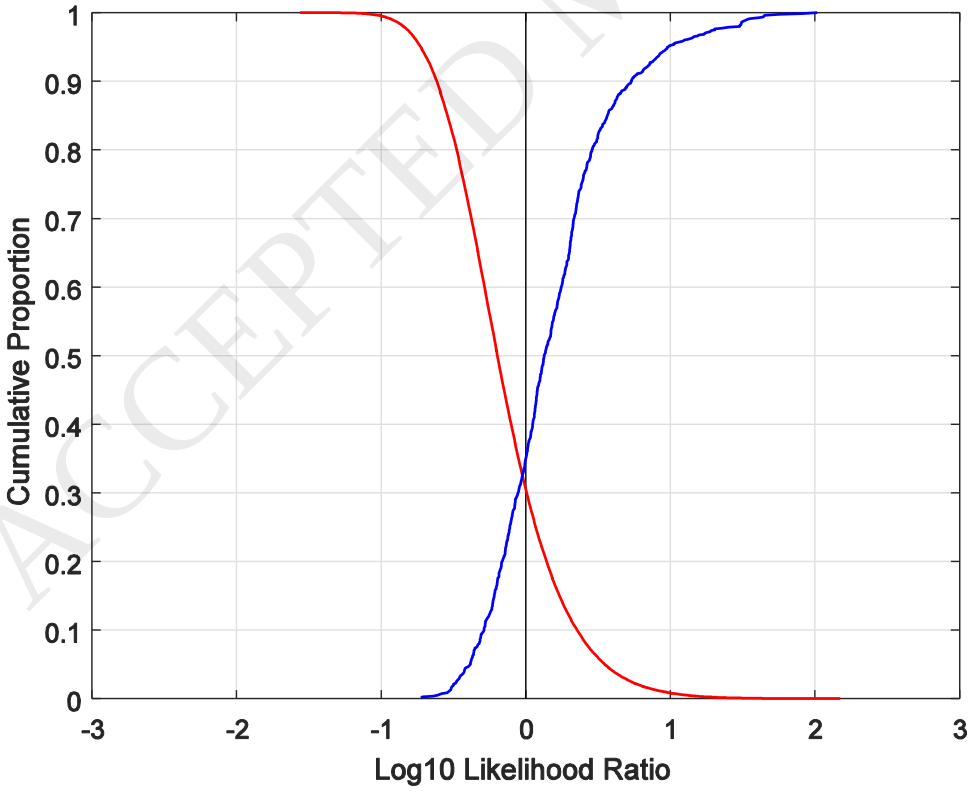
Fig. 2. Tippett plot of the results of testing the system variant with known-speaker-condition data used to train the UBM and no calibration.

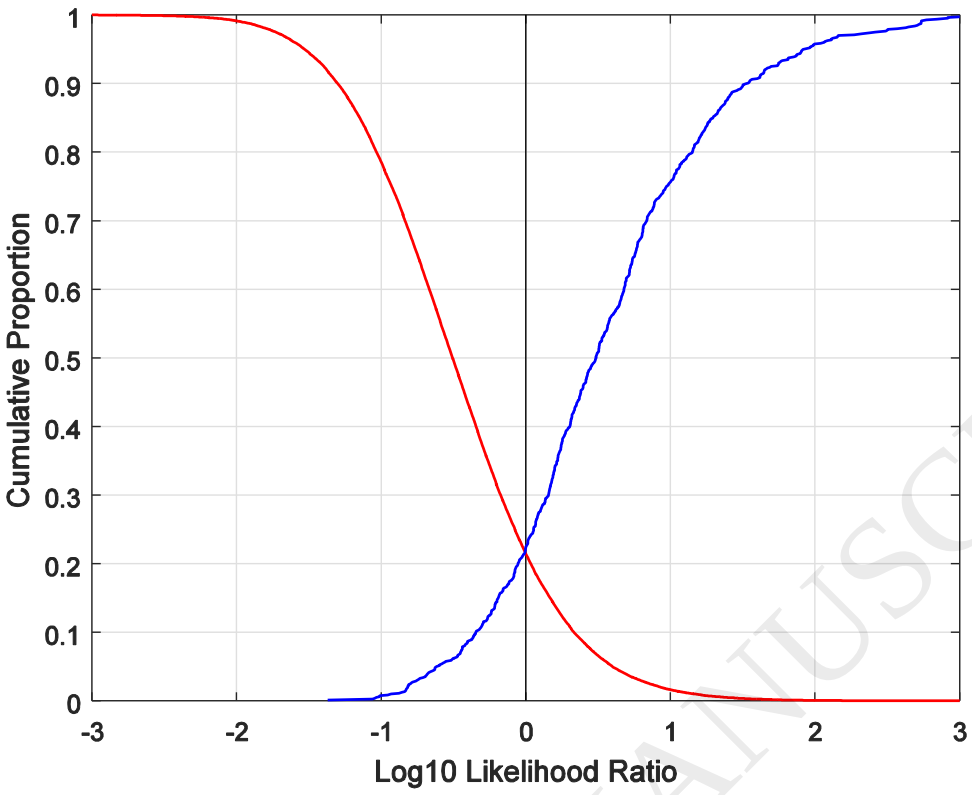
Fig. 3. Tippett plot of the results of testing the system variant with high-quality audio data used to train the UBM and with calibration trained using know-speaker-condition and questioned-speaker-condition pairs.

Fig. 4. Tippett plot of the results of testing the system variant with known-speaker-condition data used to train the UBM and with calibration trained using know-speaker-condition and questioned-speaker-condition pairs.









8 Table

Table 1. Log likelihood ratio costs for the results from each GMM-UBM system variant tested.

UBM training data:	high-quality	known-speaker-condition	high-quality	known-speaker-condition
calibration:	no	no	yes	yes
C_{llr}	0.957	0.952	0.877	0.674
C_{llr}^{min}	0.844	0.654	0.850	0.658