LR-based forensic voice comparison under severe test-data scarcity

Yuko Kinoshita[1,2] and Michael Wagner[2,3,1]

[1]Australian National University,
[2]National Centre for Biometric Studies, [3]University of Canberra
michael.wagner@ncbs.com.au, Yuko.Kinoshita@anu.edu.au

Abstract

This study sets out to find the most reliable method for log-likelihood-ratio (LLR) calculation under severe data scarcity, which is typical of forensic voice comparison casework. We compared the performances of three types of speaker modelling, namely a single Gaussian model, Gaussian Mixture Models (GMM) of different complexity, and a Multivariate Kernel Density Model (MVKD), using two and three-dimensional formant frequency feature vectors extracted from /i:/ vowels. We varied the number of tokens used in the offender dataset from 2 to 6. We find that calibration of the systems was critical for dependable evaluation with all the systems tested and that the MVKD model outperformed Gaussian models in most cases.

Index Terms: forensic voice comparison, speaker recognition, multivariate kernel density model, GMM-UBM, formant frequencies

1. Introduction

Likelihood ratio (LR) based evaluation of evidence has been a part of forensic science for a few decades now. It was first used in DNA testing, and since the beginning of this century, it has become an important part of research and practice in both automatic speaker verification [1] and forensic voice comparison [2]. There are various approaches to calculating LRs; Gaussian Mixture Model-Universal Background Model (GMM-UBM [3]) and Multivariate Kernel Density Likelihood Ratio (MVKD [4]) being used most commonly in voice comparison research. Despite their limitations, which will be further discussed below, various studies demonstrated that the LRs produced with these methods are effective in distinguishing pairs of speech samples of same-speaker origin from those of different-speaker origin (e.g. [5-10]). Automatic speaker verification, driven to a large extent by a series of NIST Speaker Recognition Evaluations, has considerably advanced the field, particularly in the problem areas of differing recording conditions and transmission channels between the speech recordings to be compared [11].

Despite this progress, however, applying these methods to real-life casework has not been straightforward. One of the reasons is the nature of the data that is available to us in real-life casework.

Firstly we have the database availability problem. Most of the research advancements come at the price of requiring many hours, if not hundreds of hours, of speech data which span those differing conditions and thereby allow the speaker recognition systems to learn how to compensate for such differences if they are encountered between pairs of speech recordings. In most cases, forensic scientists do not have access to a database that is large enough and has suitable characteristics to be used in the particular case at hand.

The recording conditions of forensic samples are also serious problems. They are usually far inferior in both quantity and acoustic quality to the datasets that are commonly used in speaker recognition research. The offender samples from crime scenes are often very brief. Also, more often than not they are compromised by background noise, such as background music, radio and television sets, car engine noise, or competing speech from other persons present. The suspect samples are typically recordings from police interviews. The acoustic quality is generally better than that of crime scene recordings, but they often contain ambient noise or acoustic artefacts of the room. In addition, the forensic scientist may have very little data from such interviews, as some suspects refuse to speak knowing that the recordings may be used against them. Furthermore, the forensic scientist is likely to be challenged by other factors such as emotional speech, speaking style mismatch, and ethnic or cultural accents in forensic samples. In short, real-life casework comes with much more complicated conditions than those in which most theoretical research work has been conducted.

This study therefore focuses on one of those adverse conditions commonly found in forensic voice comparison: scarcity of test data. We compare the performances of various LR calculation methods in a situation where not many samples are available for testing, and examine which approach is most dependable under such conditions.

Typical choices for LR calculation methods in forensic voice comparison are the modelling of the background and suspect data using single Gaussian probability density functions (pdf), GMM or MVKD. A comparative study between MVKD and GMM-UBM has already been performed and the superiority of GMM-UBM was reported in [6]. The sample sizes of the testing data in [6] were, however, 16-20 tokens per vowel, which are much greater numbers than what is available to forensic scientists in many casework situations.

LR calculation relies heavily on modelling the probability densities of within- and between-speaker differences, and it is known that small datasets cannot model the distribution as well as large datasets do. For instance, the sample size of the test data affects the quality and utility of the resulting LR, and it was found in [12] that the discrimination performance improves as the sample size increases while the calibration loss diminishes steadily. Since the calibration can be adjusted by post-processing [13], we can deduce that the sample size of the test data affects the overall performance of LR based voice comparison.

Is GMM-UBM still the best performing approach even when we have only a small amount of data for offender and suspect? How many mixtures are suitable in such case? Or does MVKD actually outperform GMM-UBM? Using the first three formants (F1, F2 and F3) of the single Australian English vowel /i:/ as the acoustic features, we examine which of these three techniques is more robust when the testing

samples are limited to the size common in casework situations.

2. Procedure

2.1. Data

This study used two databases. For the test data, we used a database of 27 Australian male speakers, which was originally built for collecting diphthongs. The speakers in this database were recorded in two sessions separated by an average of two weeks, and at least 10 days apart. This database was first studied in [14], and also used in [15-18]. As a part of the recording task, the speakers were required to spell various words. Six /i:/ vowel tokens were extracted from the sections where the speakers spelled out the words "bide" (/biː/, /ɑe/, /diː/, /iː/) and "bite" (/biː/, /ɑe/, /tiː/, /iː/). This resulted in 12 tokens altogether from each speaker (six tokens × two recording sessions). Their speaking styles were semi-spontaneous. The speakers had to spell these words from memory, as the flash cards that cued words to be spelled were shown only a few seconds and hidden before they started the spelling task. Praat [19] was used to extract F1-F4 at the midpoint of the vowel duration, although this study excluded F4 from the analysis.

The midpoint of the vowel duration was chosen as the sampling point for two reasons. First of all, it allows mechanical selection and consequently guarantees consistency of sampling across all tokens. Secondly, it is a simple way to reduce the effects of co-articulation with the adjacent segments.

F1 of /iː/ vowel is rarely used in actual casework situation, since the majority of forensic samples are recorded over the telephone and F1 is attenuated by the bandwidth of telephone transmissions. However, we decided to include F1 in this study, as the aim of the study is to examine the *relative* performances of the different LR calculation methods, and *absolute* LR values or the strength of the /iː/ vowel as evidence is not the primary interest of this study.

For the background population, we used Bernard's formant data [20], which contain F1-F3 of Australian vowels from 170 speakers. Bernard's database is not ideal for evaluating forensic casework for two reasons. Firstly it was collected in the 1960s, and since then Australian vowels have changed noticeably [21]. Also, the utterances in this database are extremely well controlled, i.e. produced in /hVd/ contexts and in isolation. They are thus in very different conditions from typical forensic speech samples and also from the testing data used in this study. However, this should not be a problem for the purpose of our study, since it does not interfere with our aim: investigating the robustness of various LR calculation methods against small sample size. We feed the same datasets to all the LR calculation methods and evaluate the performance relative to one another, but not in absolute terms.

2.2. Experiment

With each of the 27 speakers in the test dataset, all six tokens of the speaker's Session 1 data were used to build the suspect model. Session 2 data were used for building the offender model, but the number of tokens used was varied in increments of two by choosing 2, 4 and 6 tokens, respectively.

In the same-speaker comparisons, the suspect model built from Session 1 data and the offender dataset built from Session 2 data of the same speaker were compared. For the between-speaker comparisons, each of the 27 suspect models built from Session 1 data were compared to the tokens from Session 2 of the remaining 26 speakers, resulting in 27×26=702 such trials. In both types of comparisons, Bernard's formant data from 170 speakers were used to model the background population, and LRs were calculated as the ratio between the likelihood of observing the test samples given the hypothesis that the test samples are generated by the suspect model divided by the likelihood of observing the test samples given the hypothesis that the test samples are generated by the background model.

In the calculation of log likelihood ratios (LLRs), the suspect model was always built from all six tokens of Session 1. For the Gaussian models, it was represented by a mean vector (two-dimensional for F1-F2 and three-dimensional for F1-F2-F3) and a diagonal covariance matrix. For the kernels of the MVKD, mean vectors and diagonal covariances were also used. For the offender models, each speaker had three different models built from different numbers of tokens (i.e. 2, 4, and 6). The final outcomes are compared across the different techniques used.

2.2.1. Calculation of scores

Three methods were used in the LLR calculation: (1) a single multivariate Gaussian pdf; (2) four different GMM-UBMs with 2, 4, 8 and 16 mixtures, respectively, and with diagonal covariances; and (3) MVKD. Hence we evaluated a total of six different detectors.

The single Gaussian model assumes that the 2-dimensional (2D) or 3-dimensional (3D) formant vectors are distributed normally in a single cluster, an assumption that is likely to be a better approximation for low-dimensional formant vectors than for typically 12-40 dimensional MFC vectors. This model is particularly attractive when the number of samples for building the model is too small for a GMM, but can still be sufficient for training a 2D or 3D single Gaussian with diagonal covariance.

GMM allows distributions to fall into more than one cluster and, thus, provide a more realistic representation of speaker data, even when the samples are only 2D or 3D formant vectors. However, for reliable modelling, GMM requires more data than a single Gaussian. Hence it is not obvious which of the two algorithms is preferable under practical forensic casework conditions of very sparse data.

MVKD has the capacity to represent the formant data of the background population as a multimodal pdf of many kernels, each of which is a single Gaussian with a mean vector and diagonal covariance. Each kernel can be considered to be representing the distribution of formant frequencies from individuals in the background population. Each MVKD comparison, whether it is a same-speaker or different-speaker comparison, results in a single LR, which takes into account correlations between the features (e.g. between F1 and F2 of a single vowel) and is therefore considered particularly suitable for speech data where features are expected to have some correlation [22].

2.2.2. Post-processing and evaluation of the results

Prior research has shown that post-processing of the scores can improve the performance of speaker verification and forensic voice comparison considerably [17, 23, 24]. We evaluated our six detectors with an application-independent metric, the log-likelihood-ratio cost (C_{llr}), proposed by Brümmer and du Preez in [25]. Application independence

means that neither the prior probabilities nor the error costs for a specific application are given, but the detector is evaluated by estimating expected values for priors and costs by averaging over the ranges of their values. Therefore, we used this metric to evaluate relative performances of the six detectors and of three offender models built from differently-sized datasets. We also examined the effects of calibration by producing C_{llr} for both the calibrated and uncalibrated detector scores. We calibrated the resulting LLRs of our experiments using the Focal toolkit [13].

3. Results

3.1. 2-dimensional feature vectors (F1 and F2)

Tables 1 and 2 show the results obtained from 2D F1-F2 feature vectors. Each row shows the modelling of background population and suspect speech: a single Gaussian, GMM with 2, 4, 8 and 16 mixtures, and an MVKD. The three columns show the number of tokens used to create an offender model. Table 1 presents the C_{llr} produced from uncalibrated LLRs, and Table 2 presents that from the calibrated LLRs.

Firstly it is obvious that for all experimental conditions the overall cost of the system is considerably higher when the systems were not calibrated. This demonstrates clearly the value of calibrating system scores in each and every system evaluation. The calibration appears to be particularly critical for Gaussian-based systems, as they made much greater improvement on C_{llr} compared to MVKD.

Table 1. Uncalibrated C_{llr} values for 6 types of speaker modelling and 3 offender sample sizes, based on 2-dimensional F1-F2 feature vectors. The best C_{llr} value is shown in bold.

	Number of offender tokens		
Model	**2**	**4**	**6**
Normal	0.788	0.679	0.692
GMM-2	0.887	0.883	0.883
GMM-4	0.878	0.880	0.881
GMM-8	0.879	0.882	0.882
GMM-16	0.893	0.891	0.897
MVKD	***0.614***	0.647	0.657

Table 2. Calibrated C_{llr} values for 6 types of speaker modelling and 3 offender sample sizes, based on 2-dimensional F1-F2 feature vectors. The best C_{llr} value is shown in bold.

	Number of offender tokens		
Model	**2**	**4**	**6**
Normal	0.546	0.485	***0.463***
GMM-2	0.582	0.546	0.561
GMM-4	0.576	0.591	0.577
GMM-8	0.569	0.599	0.597
GMM-16	0.661	0.653	0.670
MVKD	0.518	0.528	0.507

The MVKD system generally performs better than Gaussian-based systems, regardless of the number of offender tokens. Once calibrated, the single Gaussian model (shown as "Normal" in the tables) outperformed MVKD when the offender datasets contained more than four tokens. However, on the whole, MVKD seems to be more reliable than GMM where the amount of data for testing sample was limited. We also found that the single Gaussian model is markedly better than GMMs for all offender data sizes.

3.2. 3-dimensional feature vectors (F1, F2 and F3)

Tables 3 and 4 show the uncalibrated C_{llr} and the calibrated C_{llr} , respectively, for the experimental conditions that use 3D F1-F2-F3 feature vectors. It is striking to see how poorly the single Gaussian models performed without calibration. Considering that the single Gaussian models performed better than GMM for both calibrated and uncalibrated systems with 2D feature vectors, this is noteworthy. The single Gaussian models work well only when the feature vector is very small.

With or without calibration, the MVKD systems outperformed all other systems, although the calibration significantly reduced the gap between MVKD and Gaussian-based systems. Calibration improved the C_{llr} of the Gaussian-based system much more than that of the MVKD.

Table 3. Uncalibrated C_{llr} for 6 types of speaker modelling and 3 offender sample sizes, based on 3-dimensional F1-F2-F3 feature vectors. The best C_{llr} value is shown in bold.

	Number of offender tokens		
Model	**2**	**4**	**6**
Normal	2.221	2.435	1.925
GMM-2	0.851	0.848	0.844
GMM-4	0.871	0.860	0.853
GMM-8	0.882	0.874	0.871
GMM-16	0.881	0.861	0.863
MVKD	***0.584***	0.801	0.687

Table 4. Calibrated C_{llr} for 6 types of speaker modelling and 3 offender sample sizes, based on 3-dimensional F1-F2-F3 feature vectors. The best C_{llr} value is shown in bold.

	Number of offender tokens		
Model	**2**	**4**	**6**
Normal	0.603	0.512	0.545
GMM-2	0.538	0.510	0.496
GMM-4	0.581	0.522	0.500
GMM-8	0.639	0.596	0.559
GMM-16	0.624	0.553	0.566
MVKD	0.504	0.506	***0.451***

3.3. Number of tokens in the offender datasets

With respect to the number of the tokens used as the offender dataset, we did not observe clear linear effects. For instance, in Tables 1 and 3, we observed the best performance where we used only 2 tokens in the offender dataset. Tables 2 and 4, on the other hand, showed the best performance when 6 tokens were used. We expected to have a better model (hence more accurate evaluation) with more data, so the results in Tables 1 and 3 are somewhat intriguing. We suspect that this might be accidental, caused by the particular tokens that happened to be included, since the average results can be very good or bad by accident with such a limited amount of data. Perhaps after

calibration such accidental elements were smoothed out, and more predictable results were produced. When only 2 tokens were used as the offender data, /i:/ tokens extracted from /ti:/ and /i:/ in spelling the word "bite" were used. We suspect that the results might have come out differently even without calibration, if we had tested with all the permutations of /i:/ tokens.

3D F1-F2-F3 feature vectors offer only a small advantage over 2D F1-F2 feature vectors, presumably due to the well-known fact that modelling in three dimensions requires more training data than modelling in two dimensions.

4. Conclusions and future direction

This study compared several methods of LR calculation under the constraints which we face in real-life casework scenario: testing datasets of limited samples sizes. A preceding study reported superiority of GMM to the MVKD where sufficient testing data are available [6]. However, under more forensically realistic conditions of more severe scarcity of test data, we discovered that the voice evidence could be more reliably assessed with the use of MVKD. We also found that calibration is essential to perform reliable voice comparisons regardless of the calculation methods, although the calibration seemed more critical for GMM-based systems.

This experiment needs to be extended by adding more vowels to gain further insight, as forensic scientists usually combine the results from multiple vowels.

5. Acknowledgements

We would like to thank our three anonymous reviewers for their constructive comments and suggestions. However, we are solely responsible for any remaining errors and shortcomings that may exist in this paper.

6. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech communication,* vol. 52, pp. 12-40, 2010.

[2] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science and Justice,* vol. 49, pp. 298--308, 2009.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19--41, 1// 2000.

[4] C.G.G. Aitken, and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics,* vol. 53, pp. 109--122, 2004.

[5] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech and Language,* vol. 20, pp. 331-355, 2006.

[6] 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, 2// 2011.

[7] P. J. Rose and E. Winter, "Traditional Forensic Voice Comparison with Female Formants: Gaussian mixture model and multivariate likelihood ratio analyses," in *SST2010*, Melbourne, 2010, pp. 42--45.

[8] Y. Kinoshita, S. Ishihara, and P. Rose, "Beyond the Long-term Mean: Exploring the Potential of F0 Distribution Parameters in Forensic Speaker Recognition," in *ODYSSEY 2008 - The Speaker and Language Recognition Workshop*, Stellenboch, 2008.

[9] P. J. Rose, D. Lucy, and T. Osanai, "Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical effects model: A "non-idiot's bayes" approach," in *the 10th Australian International Conference on Speech Science & Technology*, Sydney, 2004, pp. 402--407.

[10] Y. Kinoshita, S. Ishihara, and P. Rose, "Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition," *International Journal of Speech Language and the Law,* vol. 16, pp. 91-111, 2009.

[11] J. Gonzalez-Rodriguez, "Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014)," *Loquens,* vol. 1, p. e007, 2014.

[12] Y. Kinoshita and S. Ishihara, "The effect of sample size on the performance of likelihood ratio based forensic voice comparison," in *The 14th Australasian International Conference on Speech Science and Technology*, Sydney, 2012.

[13] N. Brümmer, "FoCal: Toolkit for fusion and Calibration," ed.

[14] Y. Kinoshita and T. Osanai, "Within speaker variation in diphthongal dynamics: What can we compare," in *Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand. Australia: Australasian Speech Science & Technology Association, Canberra*, 2006, pp. 112-117.

[15] G. S. Morrison, "Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/," *International Journal of Speech, Language and the Law,* vol. 15, pp. 249--266, 2008.

[16] G. S. Morrison, "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories," *Journal of the Acoustical Society of America,* pp. 2387–2397, 2009.

[17] G. S. Morrison, "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongsa)," *The Journal of the Acoustical Society of America,* vol. 125, pp. 2387-2397, 2009.

[18] G. S. Morrison and Y. Kinoshita, "Extraction of likelihood-ratio forensic evidence from the formant trajectories of diphthongs," presented at the Acoustics 2008, Paris, 2008.

[19] P. Boersma and D. Weenink, "Praat," 5.3.77 ed, 2014.

[20] J. Bernard, "Toward the acoustic specification of Australian English," *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung,* vol. 23, pp. 113-128, 1970.

[21] F. Cox and S. Palethorpe, "Reversal of short front vowel raising in Australian English," in *INTERSPEECH*, 2008, pp. 342-345.

[22] Y. Kinoshita and S. Ishihara, "Comparative performance of univariate and multivariate likelihood ratios in forensic speaker recognition," presented at the HSCnet SummerFest07, Sydney, 2007.

[23] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and Javier Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," *IEEE Transactions On Audio, Speech, and Language Processing,* vol. 15, pp. 2104-2115, September 2007.

[24] G. S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system," *Forensic science international,* vol. 208, pp. 59-65, 2011.

[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*. vol. 1, C. Müller, ed., Berlin: Springer, 2007, pp. 330--353.