Preliminary performance comparison between PCAKLR and GMM-UBM for computing the strength of speech evidence in forensic voice comparison

Hanie Mehdinezhad [1, 2], Bernard J. Guillemin [1, 2]

[1] Forensic and Biometrics Research Group (FaB), University of Auckland, New Zealand
[2] Department of Electrical and Computer Engineering, University of Auckland, New Zealand

Kmah320@aucklanduni.ac.nz,bj.guillemin@auckland.ac.nz

Abstract

A preliminary performance comparison between two probabilistic procedures for the calculation of Likelihood Ratios (LRs) in a Forensic Voice Comparison (FVC) is presented in this paper. One of these, Gaussian Mixture Model–Universal Background Model (GMM-UBM), is common in FVC. The other, Principal Component Analysis Kernel Likelihood Ratio (PCAKLR), is a relatively new procedure. Mel-Frequency Cepstral Coefficients (MFCCs) of three vowels of /aɪ/, /eɪ/ and /i:/ were the speech features used. Scores for each vowel were calibrated and fused using logistic regression. For these experiments PCAKLR is shown to outperform GMM-UBM in terms of both accuracy and reliability.

Index Terms: FVC, MFCCs, GMM-UBM, PCAKLR

1. Introduction

The Bayesian Likelihood Ratio (LR) framework is gaining increased acceptance for evaluating the strength of evidence in a Forensic Voice Comparison (FVC) [1-3]. Multivariate Kernel Density (MVKD) [4] and Gaussian Mixture Model – Universal Background Model (GMM-UBM) [5, 6] are widely used procedures for calculating LRs. The former is primarily designed for token-based analysis, while the latter is primarily designed for data-stream-based analysis [7]. Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) [8, 9], a procedure proposed by researchers at the University of Auckland, is a relatively new approach for computing LRs that is also primarily designed for token-based analysis.

Morrison compared MVKD and GMM-UBM when applied to tokenized data and reported that the later outperformed the former in terms of both accuracy and reliability [10]. Nair et al compared MVKD and PCAKLR for tokenized data and reported that for a large number of input parameters, PCAKLR outperforms MVKD in terms of accuracy [9]. To our knowledge there has not been a similar comparison study between PCAKLR and GMM-UBM. So the goal of this paper is to present a preliminary performance comparison between them for tokenized data.

The remainder of this paper is structured as follows. Section 2 provides an overview of the LR framework, followed by a brief discussion of PCAKLR and GMM-UBM. Section 3 describes our experimental procedure for comparing their performance when applied to tokenized data. The results of these experiments are presented in Section 4, followed by conclusions in Section 5.

2. Background information

2.1. Likelihood Ratio Framework

Mathematically the LR is calculated as:

$$LR = \frac{P(E|H_P)}{P(E|H_d)}.$$

$P(E|H_P)$ is the conditional probability of E (the evidence) given Hp (the prosecution hypothesis) and assesses the similarity between the suspect and offender speech samples. $P(E|H_d)$ is the conditional probability of E given H_d (the defense hypothesis) and measures the typicality of the suspect and offender speech samples to a relevant background population. LR values significantly greater than one support the prosecution hypothesis, LR values significantly less than one support the defense hypothesis, and LR values close to one provide little support either way. It is common to compute the Log-Likelihood-Ratio (LLR), where $LLR = log_{10}(LR)$, its sign indicating whether it supports the prosecution (positive) or defense (negative) and its magnitude indicating the strength of that support.

2.2. Overview of GMM-UBM and PCAKLR

2.2.1. GMM-UBM

GMM-UBM [5, 6] is common in both automatic speaker recognition and FVC. Normally it requires a large amount of data to build a single background model, namely a Universal Background Model (UBM). In order to achieve good performance, the UBM is trained on all background data pooled across speakers. The probability density function of the UBM is estimated using Gaussian Mixture Models (GMMs), with the Expectation Maximization (EM) algorithm being used to train it. The suspect model is then built by copying the UBM and adapting it towards a better fit of the suspect speech data using the Maximum a posterior (MAP) procedure. A score is then calculated as the ratio of the suspect and background probability density function values determined at the offender data points. (Note: A score is calculated using the same expression as for the LR defined above. Once it has been calibrated, it becomes an LR [11, 12])

2.2.2. PCAKLR

PCAKLR is modelled on MVKD [8, 9]. The main difference between the two is that MVKD was designed for a small number of input parameters (typically 3-4), whereas PCAKLR can handle any number of parameters. For both procedures, a normal distribution is used to model the suspect data, while a kernel density distribution is used to model the background data. The distinguishing feature of PCAKLR in

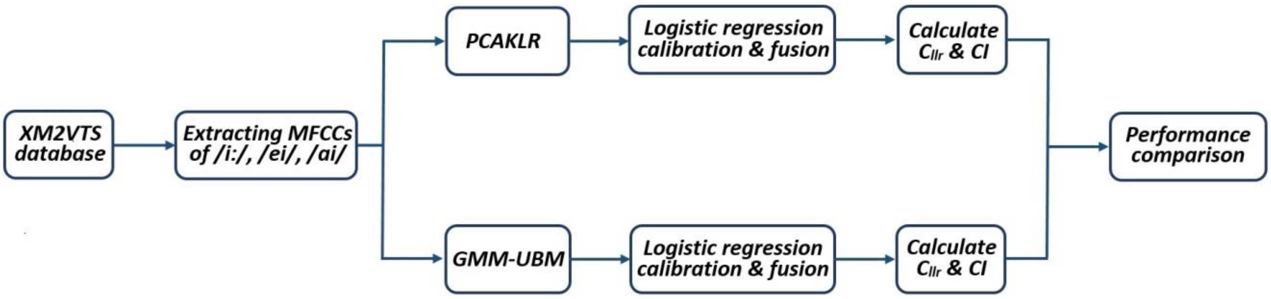
XM2VTS database
Extracting MFCCs of /i:/, /ei/, /ai/
PCAKLR
GMM-UBM
Logistic regression calibration & fusion
Logistic regression calibration & fusion
Calculate C_{llr} & CI
Calculate C_{llr} & CI
Performance comparison

Figure 1: *Experimental set-up for comparing the performance of GMM-UBM and PCAKLR*

comparison to MVKD is the manner in which it takes account of correlations between parameters. It does this by transforming the speech features into a new set of uncorrelated parameters using Principal Component Analysis (PCA). Then individual scores are computed for each of these uncorrelated sets using Univariate Kernel Density (UKD) analysis [13]. Since the transformed parameters are uncorrelated, a final score can be determined by multiplying the individual scores.

2.3. Measuring performance of a FVC / Presenting results

The results of a FVC are often presented using Tippett plots. A Tippett plot, introduced by Meuwly [14], represents the cumulative proportion of the LLR values for both same-speaker and different-speaker comparisons.

The performance of a FVC is measured by evaluating its accuracy and reliability [3]. The accuracy or validity indicates the closeness of the obtained result with the true value of the output. The Log-Likelihood Ratio Cost (C_{llr}) [15, 16] is the recommended metric for assessing this, the lower the value, the better the accuracy. The reliability or precision measures the amount of variation that could be expected in the LR values arising from such factors as the Background set used being necessarily a limited sample of the specified Background population. The Credible Interval (CI) [16] is a popular metric for evaluating this, and again, the lower this value, the better.

3. Experimental Procedure

As shown in Figure 1, using the extracted tokens of /aɪ/, /eɪ/ and /i:/, LRs were calculated using the GMM-UBM and PCAKLR procedures and their performance compared. The following sections expand upon aspects of this process.

3.1. Speech data set

The XM2VTS (Extended Multi Modal Verification for Teleservices and Security) speech database [17] was used in this research. This multi-modal database includes speech recordings digitized at 16 bits and sampled at 32 kHz. The language is English with predominantly a Southern British accent. The database contains four recording sessions of 295 subjects (156 male, 139 female) collected over a period of 4 months. Sessions were recorded at one-month intervals and during each session each speaker repeated three sentences twice. The first two sentences were random sequences of digits from zero to nine: “zero one two three four five six seven eight nine” and “five zero six nine two eight one three seven four”. The last sentence was: “Joe took Father’s green shoe bench out”.

It should be noted that the XM2VTS database contains recordings of read speech and that the level of the background noise is low. Thus from that perspective it is not forensically realistic [18]. However, in support of its use in this investigation, it does consist of a large number of speakers with a similar accent and includes multiple recordings separated by reasonable periods of time, both aspects being important in the FVC arena.

Of the 156 male speakers in this database, only 130 were used in this study. The other 26 speakers were discarded because their recordings were either less audible, or they were judged to have different accents to the rest of the speakers (see [18] for the rationale behind discarding recordings on the basis of dissimilar accent). Two diphthongs /aɪ/ and /eɪ/ and one monophthong /i:/ were extracted from these recordings from the words “nine”, “eight” and “three”, respectively. So each recording session produced four tokens of each vowel.

Mel-Frequency Cepstral Coefficients (MFCCs) are currently the most popular speech feature used in both the automatic speaker recognition and FVC arenas and have been shown to give good comparison performance [19-21], so they were used in this study also. Their good comparison performance is likely due to the fact that MFCCs are related to the perceptual parameters of the speech signal (i.e., the non-linear response of the human hearing mechanism). We have chosen to use 23 MFCCs in our experiments. This is because the speech data was down-sampled to 8 kHz (a typical value for forensic speech data acquired from landline or mobile phone networks) and at this sampling rate, a maximum of 23 MFCCs can be extracted [22].

3.2. Comparison Process

The 130 male speakers were divided into three mutually exclusive sets: 44 speakers for the Background set and 43 speakers each for the Development and Testing sets. (Note: The FVC results from the Development set are used to calibrate and fuse the results from the Testing set [11,12].) Data from three of the four recording sessions were used for the speakers in the Background set, while all four recording sessions were used for each of the speakers in the Development and Testing sets.

Table 1 shows how comparisons were undertaken, the procedure being identical for the Testing and Development sets. In respect to forming the suspect model for each comparison, the data from recording Sessions 3 and 4 were combined, giving eight tokens per vowel. For same-speaker comparisons, Sessions 1 and 2 recordings were used in turn for the offender data. With reference to Table 1, and considering same-speaker

Table 1: *FVC Comparison process*

Speakers	Same-speaker comparisons	Different-speaker comparisons
1	1-S1 vs 1-S34	1-S1 vs 2-S34, 3-S34, …, 43-S34
	1-S2 vs 1-S34	1-S2 vs 2-S34, 3-S34, …, 43-S34
		1-S3 vs 2-S34, 3-S34, …, 43-S34
2	2-S1 vs 2-S34	2-S1 vs 1-S34, 3-S34, …, 43-S34
	2-S2 vs 2-S34	2-S2 vs 1-S34, 3-S34, …, 43-S34
		2-S3 vs 1-S34, 3-S34, …, 43-S34
.	.	.
.	.	.
.	.	.
43	43-S1 vs 43-S34	43-S1 vs 1-S34, 2-S34, …, 42-S34
	43-S2 vs 43-S34	43-S2 vs 1-S34, 2-S34, …, 42-S34
		43-S3 vs 1-S34, 2-S34, …, 42-S34

comparisons for, say, Speaker 43, the two same-speaker comparisons are identified as 43-S1 vs 43-S34 and 43-S2 vs 43-S34. (Note: two same-speaker comparisons are required per speaker in order to compute the CI.) For different-speaker comparisons, Sessions 1, 2 and 3 were used in turn for the offender data. With reference to Table 1, and considering a different-speaker comparison between, say, Speaker 43 (offender) and Speaker 1 (suspect), the three comparisons are identified as 43-S1 vs 1-S34, 43-S2 vs 1-S34 and 43-S3 vs 1-S34. (Again, undertaking multiple different-speaker comparisons for the same pair of speakers is required in order to compute the CI.) With 43 speakers in each of the Testing and Development sets, this resulted in 43 same-speaker comparisons and 903 different-speaker comparisons (ignoring multiple comparisons required in order to compute the CI).

The results for individual vowels were then calibrated and fused using logistic regression [12]. The mean of LRs for the two same-speaker comparisons and the mean of LRs for the three different-speaker comparisons were used to calculate C_{llr}. CI for both same-speaker and different-speaker LRs was computed using the procedure outlined in [16].

Unlike the PCAKLR procedure, the GMM–UBM procedure has two parameters which need to be specified. First is the number of Gaussian components in the GMM. In this investigation this was varied from 8 to 16, the same values as used by Morrison [10]. Second is the number of MAP iterations in the adaptation process. We fixed this at 15, this again being the value Morrison used [10]. (Note: At the outset of our experiments the number of Gaussian components was varied between 1 and 30, and the number of MAP iterations was varied between 1 and 50. The results from these experiments confirmed that the values used in [10] are the best options.) For each vowel, the final choice of number of Gaussian components was made on the basis of lowest resulting C_{llr}, the goal being to try and ensure optimization of the FVC system to this particular data set.

4. Results

The Tippett plots in Figures 2 and 3 show the cumulative distribution of LLR values for GMM-UBM and PCAKLR, respectively. The solid blue curves in these figures are the same-speaker comparison results, and the solid red curves are the different-speaker comparison results. The dashed lines on either side of these solid curves represent the variation in a particular LR comparison result (i.e., LLR±CI). Also shown in these figures are mean C_{llr} and CI.

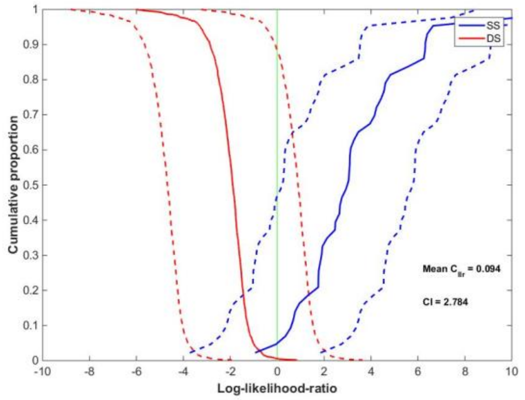

Figure 2: Tippett plot showing the performance of GMM-UBM

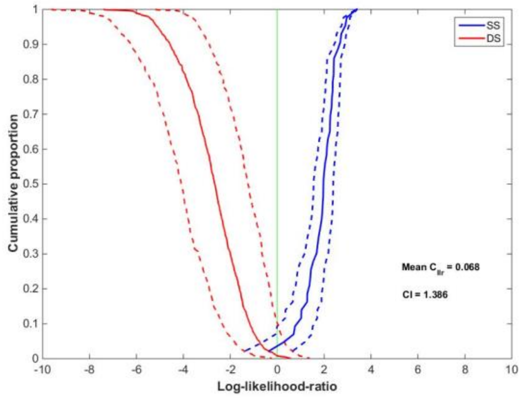

Figure 3: *Tippett plot showing the performance of PCAKLR*

It can be seen from these figures that PCAKLR has marginally outperformed GMM-UBM in terms of accuracy (C_{llr} = 0.068 compared to C_{llr} = 0.094, respectively). This seems to be mainly due to PCAKLR producing a slightly smaller number of same-speaker misclassifications[1], even though the number of different-speaker misclassifications it produced is slightly larger. In terms of reliability, PCAKLR has again outperformed GMM-UBM (CI=1.386 compared to CI=2.784, respectively), the difference now being more significant. The reason for this is not clear and further investigation is needed.

5. Conclusion

A preliminary comparison of the FVC performance of GMM-UBM and PCAKLR when applied to tokenized data has been presented in this paper. The speech feature set used was 23 MFCCs extracted from tokens of the vowels /aɪ/, /eɪ/ and /i:/ spoken by 130 male speakers from the XM2VTS speech database. Results for individual vowels were then calibrated and fused using logistic regression. In terms of both FVC accuracy and reliability, PCAKLR outperformed GMM–UBM, though the improvement in respect to accuracy was only marginal.

6. References

1. Morrison, G.S., *Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongsa).* The Journal of the Acoustical Society of America, 2009. **125**(4): p. 2387-2397.
2. Rose, P. and G. Morrison, *A response to the UK position statement on forensic speaker comparison.* The international journal of speech, language and the law, 2009. **16**(1): p. 139.
3. Morrison, G.S., *Forensic voice comparison and the paradigm shift.* Science & Justice, 2009. **49**(4): p. 298-308.
4. Aitken, C.G. and D. Lucy, *Evaluation of trace evidence in the form of multivariate data.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 2004. **53**(1): p. 109-122.
5. Reynolds, D., *Gaussian mixture models.* Encyclopedia of Biometrics, 2015: p. 827-832.
6. Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker verification using adapted Gaussian mixture models.* Digital signal processing, 2000. **10**(1): p. 19-41.
7. Jessen, M. *Comparing MVKD and GMM–UMB applied to a corpus of formant-measured segmented vowels in German*. in *International Associatio n for Forensic Phonetics and Acoustics Annual Conference (IAFPA 2014), Zurich, Switzerland*. 2014.
8. Nair, B.B., E.A. Alzqhoul, and B.J. Guillemin, *A new approach to computing likelihood ratios based on principal component analysis*, in *UNSW Forensic Speech Science Conference, Sydney, Australia. .* 2012.
9. Nair, B.B., E.A. Alzqhoul, and B.J. Guillemin. *Comparison between Mel-frequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework*. in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*. 2014.
10. Morrison, G.S., *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, 2011. **53**(2): p. 242-256.
11. Enzinger, E., G.S. Morrison, and 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.* Science & Justice, 2016. **56**(1): p. 42-57.
12. Morrison, G.S., *Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio.* Australian Journal of Forensic Sciences, 2013. **45**(2): p. 173-197.
13. Lindley, D., *A problem in forensic science.* Biometrika, 1977. **64**(2): p. 207-213.
14. Meuwly, D. and A. Drygajlo. *Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)*. in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*. 2001.
15. Brümmer, N. and J. du Preez, *Application-independent evaluation of speaker detection.* Computer Speech & Language, 2006. **20**(2): p. 230-275.
16. Morrison, G.S., *Measuring the validity and reliability of forensic likelihood-ratio systems.* Science & Justice, 2011. **51**(3): p. 91-98.
17. Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999. Citeseer.
18. Morrison, G.S., F. Ochoa, and T. Thiruvaran. *Database selection for forensic voice comparison*. in *Proceedings of Odyssey*. 2012.
19. Vergin, R., D. O'shaughnessy, and A. Farhat, *Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition.* Speech and Audio Processing, IEEE Transactions on, 1999. **7**(5): p. 525-532.
20. Ishihara, S. *The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations*. in *Australasian Language Technology Association Workshop 2013*. 2013.
21. Zhang, C., G.S. Morrison, and T. Thiruvaran. *Forensic voice comparison using Chinese/iau*. in *Proceedings of the 17th International Congress of Phonetic Sciences*. 2011.
22. Rabiner, L.R. and B. Gold, *Theory and application of digital signal processing.* Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p., 1975. **1**.