A Comparative Study of Two Statistical Modelling Approaches for Estimating Multivariate Likelihood Ratios in Forensic Voice Comparison

Shunichi Ishihara
The Australian National University
Department of Linguistics
shunichi.ishihara@anu.edu.au

Abstract

The acoustic features used in forensic voice comparison (FVC) are correlated in almost all cases. A sizeable proportion of FVC studies and casework has relied, for statistical modelling, on the multivariate kernel density likelihood ratio (MVKDLR) formula, which considers the correlations between the features and computes an overall combined likelihood ratio (LR) for the offender-suspect comparison. However, following concerns over the robustness of the MVKDLR, in particular its computational weakness and numerical instability specifically when a large number of features are employed, the principal component analysis kernel density likelihood ratio (PCAKDLR) approach was developed as an alternative. In this study, the performance of the two approaches is investigated and compared using Monte Carlo-simulated synthetic data based on the 16th-order Mel Frequency Cepstrum Coefficients extracted from the long vowel /e:/ segments of spontaneous speech uttered by 118 native Japanese male speakers. Performance is assessed in terms of validity (= accuracy) and reliability (= precision), with the log-likelihood ratio cost (C_{llr}) being used to assess validity and the 95% credible interval (95%CI) to assess reliability.

1 Introduction

In many branches of the forensic sciences, including fingerprint (Neumann et al., 2007), handwriting (Marquis et al., 2011), voice (Morrison, 2009a), DNA (Evett et al., 1993), glass fragments (Curran, 2003), earmarks and footwear marks (Evett et al., 1998), strength of evidence is widely measured using the LR framework, increasingly accepted as the standard framework for forensic inference and statistics. Calculating an LR for *voice* evidence requires, as a first step, that each individual's evidence (e.g. offender and suspect recordings) be modelled using various acoustic features (e.g. formant frequencies) that are correlated almost without exceptions. However, estimating an LR based on correlated variables is not a simple problem; it was addressed by Aitken and Lucy (2004), resulting in the development of the multivariate kernel density likelihood ratio (MVKDLR) approach. The MVKDLR has been extensively used, especially in acoustic-phonetic based forensic voice comparison (FVC) (Kinoshita et al., 2009; Morrison, 2009b), but was recently shown to be prone to instability, in particular when the number of features for modelling is too high (e.g. features ≥ 5-6). The MVKDLR formula has the propensity to collapse when some of the covariance matrices of the offender and suspect data are ill-conditioned (e.g. sparse data, large number of input features) (Nair et al., 2014, pp. 90-91). This has motivated the development of an alternative, known as the principal component analysis kernel density likelihood ratio (PCAKDLR) approach (Nair et al., 2014).

To date, FVC studies in which the PCAKDLR is used to estimate LRs are limited; thus we don't know how the PCAKDLR performs in comparison to the MVKDLR (cf. Enzinger, 2016). To address this gap in our knowledge, the current study seeks to compare the performance of the MVKDLR and PCAKDLR approaches when the number of features for modelling changes, using synthetic data generated by Monte Carlo simulations (Fishman, 1995). The outcomes (scores) of the two approaches are calibrated using the logistic-regression calibration technique proposed

Shunichi Ishihara. 2017. A Comparative Study of Two Statistical Modelling Approaches for Estimating Multivariate Likelihood Ratios in Forensic Voice Comparison. In *Proceedings of Australasian Language Technology Association Workshop*, pages 61−69.

by Brümmer and du Preez (2006). The performance of the approaches is assessed in terms of validity (= accuracy), for which the metric is the log-likelihood-ratio cost (C_{llr}) (Brümmer & du Preez, 2006), as well as reliability (= precision), for which the metric is the 95% credible interval (95%CI) (Morrison, 2011).

2 Likelihood Ratio

The likelihood ratio (LR), a measure of the quantitative strength of evidence, is a ratio of two conditional probabilities: one is the probability (p) of observed evidence (E) assuming that one hypothesis (e.g. prosecution = H_p) is true; the other is the probability of the same observed evidence assuming that the alternative hypothesis (e.g defence = H_d) is true (Robertson & Vignaux, 1995). Thus, the LR can be expressed as 1).

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \quad 1)$$

In the case of FVC, the LR will be the probability of observing the difference between the offender's and the suspect's speech samples (referred to as the evidence, E) if they had been produced by the same speaker (H_p) relative to the probability of observing the same evidence (E) if they had been from different speakers (H_d). The relative strength of the given evidence with respect to the competing hypotheses (H_p vs. H_d) is reflected in the magnitude of the LR. The more the LR deviates from unity (LR = 1; logLR = 0), the greater support for either the prosecution hypothesis (LR > 1; logLR > 0) or the defence hypothesis (LR < 1; logLR < 0).

The important point is that the LR is concerned with the probability of the evidence, given the hypothesis (either H_p or H_d), which is the province of forensic scientists, while the trier-of-fact is concerned with the probability of the hypothesis, given the evidence. That is, the ultimate decision as to whether the suspect is guilty or not does not lie with the forensic expert, but with the court. The role of the forensic scientist is to estimate the strength of evidence (= LR) with a view to help the trier-of-fact make a final decision (Morrison, 2009a, p. 229).

3 Database, target segment, and speakers

In this study, monologues from the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000) are used for FVC experiments. The recordings are 10-25 minutes long.

For this study, it was decided to target fillers. Fillers are sounds or words (e.g. *um*, *you know*, *like* in English) uttered by a speaker to signal that he/she is thinking or hesitating. The filler /e:/ and the /e:/ segment of the filler /e:to:/ were chosen because i) they are two of the most frequently used fillers in Japanese (many monologues contain at least ten of them) (Ishihara, 2010), ii) the vowel /e/ reportedly has the strongest speaker-discriminatory power out of the five Japanese vowels /i,e,a,o,u/ (Kinoshita, 2001), and iii) the segment /e:/ is significantly long so that it is easy to extract stable spectral features from this segment. It is also considered that fillers are uttered unconsciously or semiconsciously by the speaker and carry no lexical meaning. They are thus not likely to be affected by the pragmatic focus of the utterance.

For the experiments, speakers were selected from the CSJ based on five criteria: i) availability of two non-contemporaneous recordings per speaker (n.b. suspect and offender recordings are non-contemporaneous in real cases), ii) high spontaneity of speech (e.g. not reading), iii) exclusive use of standard modern Japanese, iv) presence of at least ten /e:/ segments, and v) availability of complete annotation of the data. As the researchers had real casework in mind, only male speakers were chosen for experiments. This is because males are more likely to commit crimes than females (Kanazawa & Still, 2000). The five criteria combined resulted in 236 recordings (118 speakers x 2 non-contemporaneous recordings), all of which were used in our experiments.

The 118 speakers were divided into three mutually exclusive sub-databases: the test database (40 speakers), the background database (39 speakers) and the development database (39 speakers). Each speaker in these databases has two recordings that are non-contemporaneous. The first ten /e:/ segments were annotated in each recording. Thus, for example, there are 800 annotated /e:/ segments in the test database (= 40 speakers x 2 sessions x 10 segments). Data sparsity is a common issue in FVC. Ten samples for each recording can be judged as a realistic setting. All statistics required for conducting Monte Carlo simulations were calculated using these databases.

The speaker comparisons derived from the test database were used to assess the performance of the FVC system. The background database was

used as a background reference population, and the development database was for obtaining the logistic-regression weight, which was used to calibrate the scores of the test database (refer to §4.5 for a detailed explanation of calibration).

4 Experiments

4.1 Features

We used 16 Mel Frequency Cepstrum Coefficients (MFCC) in the experiments as feature vectors. MFCC is a standard spectral feature used in many voice-related applications, including automatic speaker recognition. All original speech samples were downsampled to 16kHz before MFCC values were extracted from the mid-duration-point of the target segment /e:/ with a 20 ms wide hamming window.

4.2 General experimental design

There are two types of tests for FVC. One relies on so-called *Same Speaker Comparisons* (SS comparisons), where two speech samples produced by the same speakers are compared. They are expected to receive LR values > 1 given the same-origins. The other type of test relies on *Different Speaker Comparisons* (DS comparisons), where two speech samples produced by different speakers are compared. They are expected to receive LR values < 1 given the different-origins.

For example, the 40 speakers of the test database enable us to undertake 40 SS comparisons and 1560 (= $_{50}C_2$ x 2) independent (e.g. not-overlapping) DS comparisons. Theoretically speaking, origin being identical, the 40 SS comparisons should receive an LR > 1 ($\log_{10}LR > 0$); on the other hand, origin being different, the 1560 DS comparisons should receive an LR < 1 (or $\log_{10}LR < 0$).

4.3 Likelihood ratio calculation

MVKDLR Approach

The MVKDLR formula computes a single LR from multiple variables (e.g. 16th-order MFCC), considering the correlations among them (Aitken & Lucy, 2004).

The numerator of the MVKDLR formula calculates the probability of evidence, which is the difference between the offender and suspect speech samples, when it is assumed that both samples have the same origin (in other words, that the persecution hypothesis H_p is true). For this calculation, the feature vectors of the offender and suspect samples and the within-speaker variance, which is given in the form of a variance/covariance matrix, are needed. The same feature vectors of the offender and suspect samples and the between-speaker variance are used in the denominator of the formula to estimate the probability of getting the same evidence when it is assumed that they have different origins (i.e. that the defence hypothesis H_d is true). These within- and between-speaker variances are estimated from the background database. The MVKDLR formula assumes normality for within-speaker variance while it uses a kernel-density model for between-speaker variance.

In the MVKDLR formula, the covariance matrices for offender and suspect data are used extensively. The inverses of these matrices are required at some stages in the process, and there are also some instances of these inverted matrices being re-inverted. All of these processes contribute to the decorrelation of the original features and the equalisation of their contribution. However, in return, the MVKDLR formula has the propensity to collapse when some of the covariance matrices of the offender and suspect data are ill-conditioned due to, for example, sparse data and large input parameters.

PCAKDLR Approach

In the PCAKDLR approach, in particular when high-dimensional features are used, the issue of the instability described for the MVKDLR is handled by decorrelating the features through principal component analysis (PCA), and then estimating LRs as the product of the univariate LRs of the resultant uncorrelated features. Thus, PCA is merely used to decorrelate the features (not to reduce feature dimensionality). With the resultant orthogonal features, a univariate LR was estimated separately for each feature using the modified kernel density model (Nair et al., 2014, pp. 88-90); the independent LRs were multiplied to generate an overall LR.

4.4 Repeated experiments using Monte Carlo simulations

As explained earlier, each speaker has two sets of ten /e:/ segments, and 16 MFCC values were extracted from each of them. That is, each session of each speaker can be modelled maximally with ten sets of 16th-order feature vectors. From these ten sets of vectors for each session of each speaker, we also obtained the basic statistics (the mean vector μ and variance/covariance matrix ε) needed

for the Monte Carlo simulations. In this study, we randomly generated, for each session of each speaker, ten feature vectors, each of which consists of 16 MFCC values. We repeated this procedure 150 times using the normal distribution function modelled with the basic statistics.

Figure 1 is an example of the Monte Carlo simulation showing 150 randomly generated first two MFCC values (c1 and c2) from the normal distribution function based on the statistics (μ and ε) obtained from the first session of the first speaker in the test database.

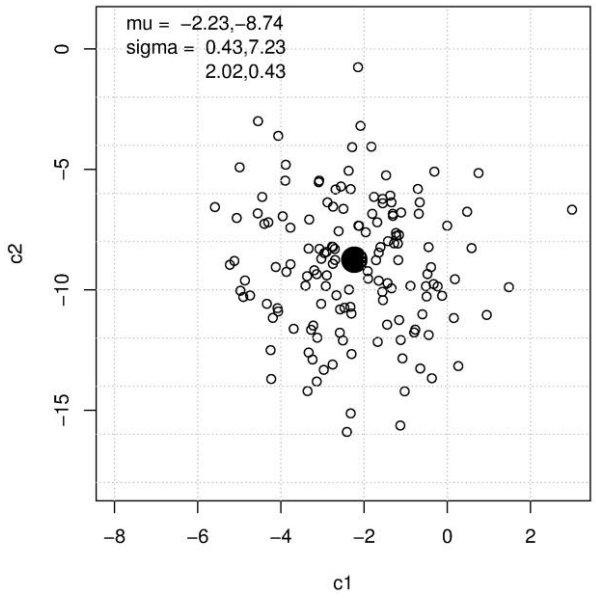

Figure 1: 150 randomly generated values (c1 and c2) from the statistics (μ and ε) obtained from the first session of the first speaker in the test database (only the first and second MFCC). The filled black circle = μ.

Experiments were repeatedly conducted using randomly generated synthetic feature vectors with different dimensions (= {2,4,6,8,10,12,14,16}). For example, a feature dimension of {2} means that the first two MFCC values were used for experiments, and a feature dimension of {14} means that the first 14 MFCC were used.

4.5 Calibration

A logistic-regression calibration (Brümmer & du Preez, 2006) was applied to the outputs (scores) of the MVKDLR and PCAKDLR approaches. Given two sets of scores derived from the SS and DS comparisons and a decision boundary, calibration is a normalisation procedure involving linear monotonic shifting and scaling of the scores relative to the decision boundary, so as to minimise a cost function, resulting in LRs. The FoCal toolkit [1] was used for the logistic-regression calibration in this study (Brümmer & du Preez, 2006). The logistic-regression weight was obtained from the development database.

4.6 Evaluation of performance: validity and reliability

The performance of the LR-based FVC system needs to be assessed in terms of its validity (= accuracy) and reliability (= precision). To explain the concepts of validity and reliability, we will look at an example. Let us imagine we have speech samples collected from two speakers at four different sessions denoted as S1.1, S1.2, S1.3, S1.4, S2.1, S2.2, S2.3 and S2.4, where S = speaker, and 1, 2, 3 and 4 = the first, second, third and fourth sessions (e.g. S1.1 refers to the first session recording collected from (S)peaker1, and S1.4 to the fourth session from that same speaker). From these speech samples, four independent (not overlapping) DS comparisons are possible: S1.1 vs. S2.1, S1.2 vs. S2.2, S1.3 vs. S2.3 and S1.4 vs. S2.4. Let us further suppose that we conducted two separate FVC tests using two different systems (Systems 1 and 2), and that we obtained the $\log_{10}$LRs given in Table 1 for these four DS comparisons.

DS comparison	System 1	System 2
S1.1 vs. S2.1	-8.3	-5.1
S1.2 vs. S2.2	-7.9	-1.2
S1.3 vs. S2.3	-8.0	-3.1
S1.4 vs. S2.4	-8.2	-0.1

Table 1: Example $\log_{10}$LRs explaining the concepts of validity and reliability.

Since the comparisons given in Table 1 are all DS comparisons, the desired $\log_{10}$LR value needs to be lower than 0, and the greater the negative $\log_{10}$LR value is, the better the system is, as it more strongly supports the correct hypothesis. For both Systems 1 and 2, all of the comparisons received $\log_{10}\text{LR} < 0$. That is, all of these $\log_{10}$LR values correctly single out the defence hypothesis. However, System 1 performs better than System 2 in that its $\log_{10}$LR values are further away from unity ($\log_{10}\text{LR} = 0$) than the $\log_{10}$LR values of System 2. This means that the $\log_{10}$LR values estimated by System 1 provide greater support for the correct hypothesis than System 2. Thus, it can be said that the validity (=

[1] https://sites.google.com/site/nikobrummer/focal

accuracy) of System 1 is higher than that of System 2. This is the basic concept of validity.

In this study, the log-likelihood-ratio cost (C_{llr}), which is a gradient metric based on LR, was used as the metric for validity. The calculation of C_{llr} is given in 2) (Brümmer & du Preez, 2006).

$$C_{llr}=\frac{1}{2}\left(\begin{array}{l}\frac{1}{N_{H_p}}\sum_{i\ for H_p=true}^{N_{H_p}}\log_2\left(1+\frac{1}{LR_i}\right)+\\ \frac{1}{N_{H_d}}\sum_{j\ for H_d=true}^{N_{H_d}}\log_2\left(1+LR_j\right)\end{array}\right) \qquad 2)$$

In 2), N_{Hp} and N_{Hd} are the number of SS and DS comparisons, and LR_i and LR_j are the linear LRs derived from the SS and DS comparisons, respectively. Given the same origins, all the SS comparisons should produce LRs greater than 1, and given the different origins, the DS comparisons should produce LRs less than 1. C_{llr} takes into account the magnitude of derived LR values, and assigns them appropriate penalties. In C_{llr}, LRs that support counter-factual hypotheses or, in other words, contrary-to-fact LRs (LR < 1 for SS comparisons and LR > 1 for DS comparisons) are heavily penalised and the magnitude of the penalty is proportional to how much the LRs deviate from unity. The lower the C_{llr} value, the better the performance.

The C_{llr} measures the overall performance of a system in terms of validity based on a cost function in which there are two main components of loss: namely discrimination loss (C_{llr}^{min}) and calibration loss (C_{llr}^{cal}) (Brümmer & du Preez, 2006). The former is obtained after the application of the so-called pooled-adjacent-violators (PAV) transformation – an optimal non-parametric calibration procedure. The latter is obtained by subtracting the former from the C_{llr}. In this study, besides C_{llr}, C_{llr}^{min} and C_{llr}^{cal} are also referred to. Once again, the FoCal toolkit[1] was used in this study for calculating C_{llr} (including both C_{llr}^{min} and C_{llr}^{cal}) (Brümmer & du Preez, 2006).

Let us now move to the concept of reliability. All of the DS comparisons given in Table 1 are comparisons of the same speaker pair (S1 vs. S2). Thus, it can be expected that the LR values obtained for these four DS comparisons should be similar as they are comparing the same speaker pair. However, the $\log_{10}$LR values based on System 1 are closer to each other (-8.3, -7.9, -8.0 and -8.2) than those based on System 2 (-5.1, -1.2, -3.1 and -0.1). In other words, the reliability (= precision) of System 1 is higher than that of System 2. This is the basic concept of reliability.

As a metric of reliability, we used 95%credible intervals, the Bayesian analogue of frequentist confidence intervals (Morrison, 2011). In this study, we calculated 95% credible intervals (95%CI) in the parametric manner based on the deviation-from-mean values collected from all of the DS comparison pairs. For example, 95%CI = 1.23 and $\log_{10}$LR = 2 means, for this particular comparison, that it is 95% certain that $\log_{10}LR \geq 0.77$ (= 2-1.23) and $\log_{10}LR \leq 3.23$ (= 2+1.23). The smaller the 95%CI, the better the reliability. The 95%CI is obtainable only from the DS comparisons in the present study.

5 Experiment with Original Data

Before presenting the results of the experiments using synthetic data, we conducted experiments using the full 16th-order MFCC values from the original databases with the two different approaches. The results of these experiments are given as Tippett plots in Figure 2 with the C_{llr} and 95%CI values. Figure 2a is for the MVKDLR and Figure 2b is for the PCAKDLR. In these Tippett plots, the solid black curve indicates the cumulative proportion of the SS comparison $\log_{10}$LRs (40) that are equal or smaller than the value indicated on the x-axis, and the solid grey curve indicates the cumulative proportion of the DS comparison $\log_{10}$LRs (1560) that are equal or greater than the value indicated on the x-axis. Tippett plots graphically show how strongly the derived LRs not only support the correct hypothesis but also misleadingly support the contrary-to-fact hypothesis. In Figure 2, the $\log_{10}$LRs for the DS comparisons are plotted together with ±95%CI band.

In terms of validity, the MVKDLR (C_{llr} = 0.396) marginally outperforms the PCAKDLR (C_{llr} = 0.418), but in terms of reliability, the PCAKDLR (95%CI = 3.536) outperforms the MVKDLR (95%CI = 4.026). As far as the Tippett plots are concerned, it can be seen from Figure 2a and Figure 2b that the magnitude of the derived LRs is very similar and comparable between the MVKDLR and PCAKDLR approaches. In terms of the discrimination (C_{llr}^{min}) and calibration (C_{llr}^{cal}) losses, although the MVKDLR (C_{llr}^{min} = 0.253 and C_{llr}^{cal} = 0.143) is slightly better than the PCAKDLR (C_{llr}^{min} = 0.267 and C_{llr}^{cal} = 0.151) in

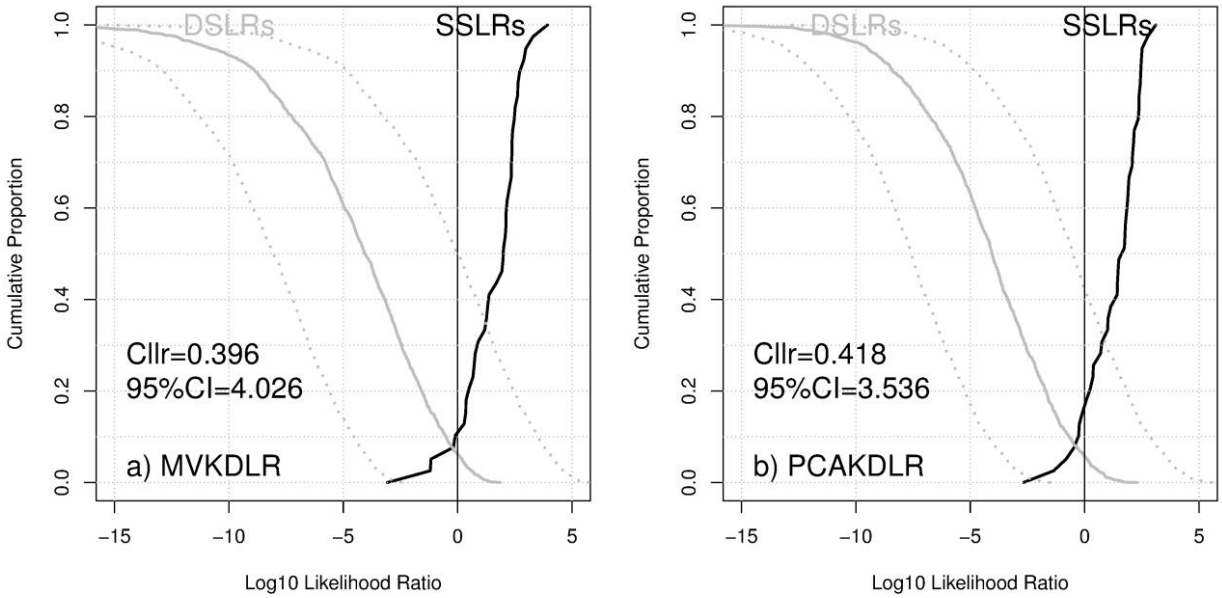

Figure 2: Tippett plots showing the magnitude of the derived LRs plotted separately for the SS (black) and DS (grey) comparisons. ±95%CI bands (grey dotted lines) are superimposed on the DS LRs. Panel a = MVKDLR and Panel b = PCAKDLR. C_{llr} value was calculated from the calibrated LRs and 95%CI value was calculated only for the calibrated DS LRs.

metric value, they are virtually the same and therefore comparable.

6 Experimental Results and Discussions

It was shown in §5 that, in terms of the C_{llr} (including both C_{llr}^{min} and C_{llr}^{cal}), the MVKDLR performed marginally better than the PCAKDLR (they performed equally well in a practical sense), but that the PCAKDLR outperformed the MVKDLR in terms of the 95%CI. It will be investigated in this section whether the observation made in §5 is a general observation that retains its validity when synthetic data are used. It will also be investigated how the number of features affects the performance of the two different approaches because the MVKDLR reportedly has an issue of instability when high-dimensional features are used.

Before the results of the experiments are displayed, it needs to be pointed out that the C_{llr} (MVKDLR = 0.396 and PCAKDLR = 0.418) and 95%CI values (MVKDLR = 4.026 and PCAKDLR = 3.536) with the 16th-order MFCC feature vector of the original data, which were given in §5, are similar to the mean C_{llr} (MVKDLR = 0.439 and PCAKDLR = 0.465) and 95%CI values (MVKDLR = 0.3348 and PCAKDLR = 2.689) of the 150 simulations with the synthetic 16th-order MFCC feature vector. This suggests the appropriateness of the Monte Carlo simulation.

In Figure 3, the mean C_{llr} (Panel a), C_{llr}^{min} (Panel b) and C_{llr}^{cal} (Panel c) values of the 150 simulations are plotted for the different feature numbers (= {2,4,6,8,10,12,14,16}) against the mean 95%CI values (y-axis), but separately for the MVKDLR (filled circles) and PCAKDLR (empty circles) approaches. The numerical information of Figure 3 is given in Table 2.

It can be seen from Figure 3a that the overall performance of the MVKDLR in terms of validity (C_{llr}) improves as the number of features increases in that there is substantial improvement when moving from two to four features (from C_{llr} = 0.939 for {2} to C_{llr} = 0.674 for {4}), after which the improvement still continues, yet to a substantially lesser degree. The general trend of the reliability (95%CI) for the MVKDLR is that it deteriorates as the feature number increases. Thus, the trade-off between validity and reliability is fairly clearly observable in the case of the MVKDLR. Although, in terms of validity, the PCAKDLR shows a more or less similar trend compared to the MVKDLR, some exceptions (e.g. the feature number = {6,8}) can also be observed in that the inclusion of additional features did not contribute to an improvement in validity (from C_{llr} = 0.712 for {6} to C_{llr} = 0.736 for {8}). These exceptions

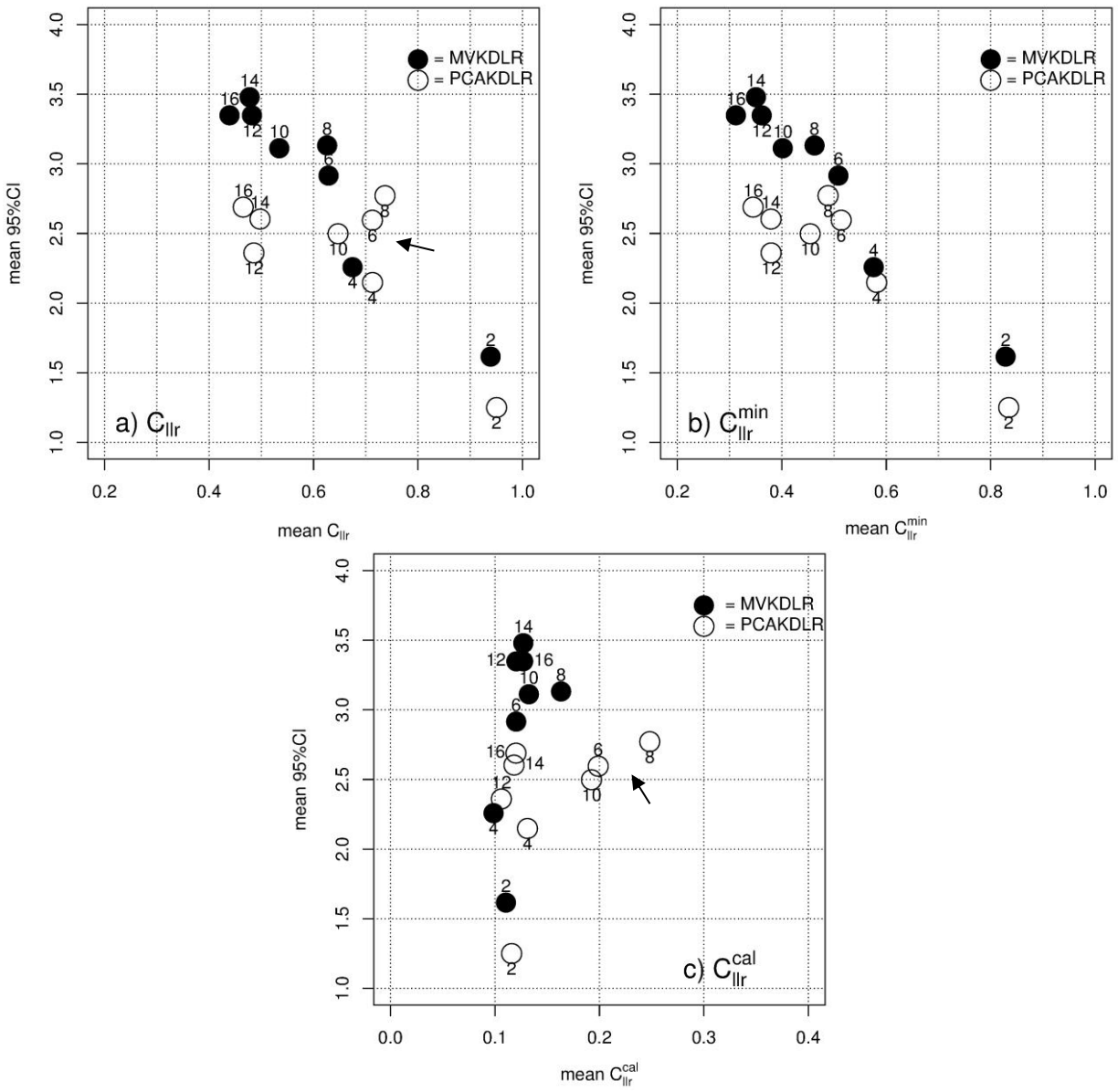

Figure 3: Mean C_{llr} (Panel a), C_{llr}^{min} (b) and C_{llr}^{cal} (c) values (x-axis) plotted against mean 95%CI values (y-axis). Filled and empty circles = MVKDLR and PCAKDLR, respectively. The number attached to each circle indicates the dimension of the feature vector. Note that the range of the x-axis scale is narrower for Panel c) than for Panels a) and b).

will be further investigated below with reference to calibration loss (C_{llr}^{cal}).

The main difference between the two approaches that can be observed from Figure 3a is that the inverse-correlation identified between validity and reliability for the MVKDLR is *not* clearly observable in the case of the PCAKDLR; the 95%CI values stay relatively constant around a 95%CI = 2.5 when six or more features are used for the experiments. As a result, although the PCAKDLR is constantly better in reliability (95%CI) than the MVKDLR, reliability is considerably better for the former than for the latter, in particular when the feature number is large (feature number ≥ 5-6). The observations derived from Figure 3a above more or less coincide with Enzinger (2016) finding that the MVKDLR is superior to the PCAKDLR in terms of validity, but inferior in terms of reliability.

Close observation of Figure 3b, which plots the discriminability of the system (C_{llr}^{min}) against its reliability shows an even clearer difference between the MVKDLR and PCAKDLR approaches described on the basis of Figure 3a in that the 95%CI values seem to hit a ceiling with six features or more for the PCAKDLR, while the 95%CI value continues to increase for the MVKDLR as a function of the feature number.

Metrics	Approaches	2	4	6	8	10	12	14	16
C_{llr}	MVKDLR	0.939	0.674	0.628	0.626	0.534	0.482	0.477	0.439
	PCAKDLR	0.950	0.712	0.712	0.736	0.646	0.485	0.497	0.465
C_{llr}^{min}	MVKDLR	0.828	0.576	0.508	0.462	0.401	0.361	0.350	0.312
	PCAKDLR	0.834	0.581	0.513	0.488	0.454	0.379	0.379	0.345
C_{llr}^{cal}	MVKDLR	0.110	0.098	0.120	0.163	0.132	0.120	0.127	0.127
	PCAKDLR	0.115	0.131	0.198	0.248	0.192	0.106	0.118	0.120
95%CI	MVKDLR	1.615	2.258	2.915	3.132	3.112	3.348	3.478	3.348
	PCAKDLR	1.250	2.148	2.595	2.771	2.498	2.360	2.603	2.689

Table 2: Numerical information of Figure 3. 2~16 = number of features

Figure 3b shows that i) system discriminability improves as a function of the number of features both for the MVKDLR and PCAKDLR approaches, ii) discriminability is marginally but constantly better for the MVKDLR, and iii) reliability is constantly better for the PCAKDLR; the former is considerably better than the latter when the feature number is large (feature number ≥ 6).

From Figure 3c, it can be seen that calibration performance is fairly constant and comparable (ca. C_{llr}^{min} = 0.125) between the MVKDLR and PCAKDLR approaches across different feature numbers, except for feature number = {6,8,10} for the PCAKDLR, which is indicated by an arrow in Figure 3c. The poor performance in calibration for feature numbers = {6,8,10} of the PCAKDLR contributes to the overall poor C_{llr} values of the PCAKDLR for the same feature numbers, which can be clearly seen in Figure 3a (as indicated by the arrow). As a result, the C_{llr} values of feature numbers = {6,8,10} are fairly better for the MVKDLR than for the PCAKDLR, while the former approach is only marginally better than the latter for the other feature numbers. However, it is not clear at this stage whether these poor calibrations are due to the PCAKDLR approach or other intrinsic or extrinsic reasons.

It has been reported in some studies that validity and reliability are often (but not always) negatively correlated (Frost, 2013; Ishihara, 2017; Morrison, 2011). That is, the better performance of the PCAKDLR in reliability may be merely due to the trade-off between validity and reliability because the MVKDLR performs better than the PCAKDLR in terms of validity. Although the effect of the trade-off should not be neglected, it is true that the PCAKDLR is substantially better in reliability than the MVKDLR, while the MVKDLR only marginally performed better than the PCAKDLR in terms of discriminability. The general trend for the 95%CI value to continue to increase as the feature number increases is another aspect of the MVKDLR that the PCAKDLR does not exhibit; the 95%CI values become saturated with a feature number ≥ 6. Thus, it is a sensible conclusion that the PCAKDLR is better than the MVKDLR in terms of reliability.

7 Conclusions and Future Directions

The outcomes of the experiments with the simulated data demonstrate some general characteristics of the PCAKDLR approach as compared to the MVKDLR approach: i) the PCAKDLR approach marginally underperforms the MVKDLR approach in terms of discriminability (C_{llr}^{min}), ii) the PCAKDLR approach performs constantly better than the MVKDLR in terms of reliability, and iii) a substantial difference in reliability performance can be observed in particular when the feature number is six or more. In some cases, the MVKDLR performed noticeably (not marginally) better than the PCAKDLR (e.g. feature number = {6,8,10}) with respect to C_{llr}, but it was pointed out that this is due to the poor calibration performance of the PCAKDLR. However, it is not clear whether these poor calibrations are indeed caused by the PCAKDLR or by other unrelated factors.

In the current study, the maximum number of /e:/ tokens, which is ten, was used to model each session of each speaker. It would be interesting to explore how a different number of tokens for modelling will influence the performance of the MVKDLR and PCAKDLR approaches because, in real cases, one is less likely to have many comparable tokens for modelling.

Acknowledgments

The author thanks the three reviewers for their valuable comments.

References

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 53*(1), 109-122.

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language, 20*(2-3), 230-275.

Curran, J. M. (2003). The statistical interpretation of forensic glass evidence. *International Statistical Review, 71*(3), 497-520.

Enzinger, E. (2016). Likelihood ratio calculation in acoustic-phonetic forensic voice comparison: Comparison of three statistical modelling approaches. *Proceedings of the Interspeech 2016*, 535-539.

Evett, I. W., Lambert, J. A., & Buckleton, J. S. (1998). A Bayesian approach to interpreting footwear marks in forensic casework. *Science & Justice, 38*(4), 241-247.

Evett, I. W., Scranage, J., & Pinchin, R. (1993). An illustration of the advantages of efficient statistical-methods for RFLP analysis in forensic-science. *American Journal of Human Genetics, 52*(3), 498-505.

Fishman, G. S. (1995). *Monte Carlo: Concepts, Algorithms, and Applications*. New York: Springer.

Frost, D. (2013). *Likelihood Ratio-based Forensic Voice Comparison on L2 Speakers: A Case of Hong Kong Male Production of English Vowels*. (Unpublished Honours thesis), The Australian National University, Canberra.

Ishihara, S. (2010). Variability and consistency in the idiosyncratic selection of fillers in Japanese monologues: Gender differences. *Proceedings of the Australasian Language Technology Association Workshop 2010*, 9-17.

Ishihara, S. (2017). Strength of forensic text comparison evidence from stylometric features: A multivariate likelihood ratio-based analysis. *The International Journal of Speech, Language and the Law, 24*(1), 67-98.

Kanazawa, S., & Still, M. C. (2000). Why men commit crimes (and why they desist). *Sociological Theory, 18*(3), 434-447.

Kinoshita, Y. (2001). *Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants*. (Unpublished PhD thesis), The Australian National University, Canberra.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech Language and the Law, 16*(1), 91-111.

Maekawa, K., Koiso, H., Furui, S., & Isahara, H. (2000). Spontaneous speech corpus of Japanese. *Proceedings of the 2nd International Conference of Language Resources and Evaluation*, 947-952.

Marquis, R., Bozza, S., Schmittbuhl, M., & Taroni, F. (2011). Handwriting evidence evaluation based on the shape of characters: Application of multivariate likelihood ratios. *Journal of forensic sciences, 56*(Supplement 1), S238-242.

Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science & Justice, 49*(4), 298-308.

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America, 125*(4), 2387-2397.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice, 51*(3), 91-98.

Nair, B., Alzqhoul, E., & Guillemin, B. J. (2014). Determination of likelihood ratios for forensic voice comparison using principal component analysis. *International Journal of Speech Language and the Law, 21*(1), 83-112.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of forensic sciences, 52*(1), 54-64.

Robertson, B., & Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: John Wiley.