The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations

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

Abstract

This study is an investigation into the effect of sample size on a likelihood ratio (LR) based forensic voice comparison (FVC) system. In particular, we looked into how the offender and suspect sample size (or the within-speaker sample size) would affect the performance of the FVC system, using spectral feature vectors extracted from spontaneous Japanese speech. For this purpose, we repeatedly conducted Monte Carlo method based experiments with different sample size, using the statistics obtained from these feature vectors. LRs were estimated using the multivariate kernel density LR formula developed by Aitken and Lucy (2004). The derived LRs were calibrated using the logistic-regression calibration technique proposed by Brümmer and du Preez (2006). The performance of the FVC system was assessed in terms of the log-likelihood-ratio cost (C_{llr}) and the 95% credible interval (CI), which are the metrics of validity and reliability, respectively. We will demonstrate in this paper that 1) the validity of the system notably improves when up to six tokens are included in modelling a speaker session, and 2) the system performance converges with the relative small token number (four) in the background database, regardless of the token numbers in the test and development databases.

1 Introduction

It is well known and accepted that statistical accuracy relies on having a sufficient amount of data. However, in typical forensic voice comparison (FVC) casework, the crime scene recording is often short and contains background noise, which limits the choice of segments that experts can use for the comparison. For example, the word *yes* is one of the most commonly used segments in FVC. However, the number of *yes* tokens we can extract from the offender sample to build his/her model really depends on the recording condition, something that forensic caseworkers cannot control. Thus, we need to know how the performance of an FVC system is influenced by sample size.

The current study employs the Likelihood Ratio (LR) framework, which has been advocated as the logically and legally correct way of analysing and presenting forensic evidence, in the major textbooks on the evaluation of forensic evidence (e.g. Robertson & Vignaux 1995), and by forensic statisticians (e.g. Aitken & Stoney 1991, Aitken & Taroni 2004), and is the standard framework in DNA comparison science. Emulating DNA forensic science, many fields of forensic sciences, such as fingerprint (Neumann et al. 2007), handwriting (Bozza et al. 2008), voice (Morrison 2009) and so on, started adopting the LR framework to quantify evidential strength (= LR).

In order to calculate an LR, we need three sets of speech samples: a set of questioned samples (offender's samples); a set of known samples (suspect's samples); and the background or reference samples. This is because an LR is a ratio of similarity to typicality, which quantifies how similar/different the questioned and the known samples are, and then evaluates that similarity/difference in terms of typicality/atypicality against the relevant background population (i.e. reference samples). Some investigations have been made on how factors such as the size and linguistic compatibility of the background population data can influence LR-based FVC (Kinoshita & Norris 2010, Ishihara & Kinoshita 2008, Kinoshita et al. 2009). Ishihara and Ki-

Shunichi Ishihara. 2013. The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations. In *Proceedings of Australasian Language Technology Association Workshop*, pages 25−33.

noshita (2008), for example, investigated how many speakers are ideally required in the background population data in order to reliably evaluate speech evidence in FVC.

However, to the best of our knowledge, studies focusing on the sample size of the offender and suspect data are conspicuously sparse. Needless to say, the sample size of the offender and suspect data – for example, the number of *yes* tokens we can use in order to build the offender's and suspect's models – has a great affect on the performance of FVC systems.

Thus, this study investigated how the offender and suspect sample sizes (or within-speaker sample size) would influence the performance of an FVC system by employing Monte Carlo simulations (Fishman 1995). In order to answer this question, two experiments: Experiments 1 and 2, were conducted. Detailed explanations of these two experiments are given §4.4.

LRs were estimated using Aitken and Lucy's (2004) MVLR formula (see §4.3). The derived LRs were calibrated using the logistic-regression calibration technique proposed by Brümmer and du Preez (2006) (see §4.5). The performance of the FVC system was assessed in terms of the log-likelihood-ratio cost (C_{llr}) (Brümmer & du Preez 2006) and the 95% credible interval (CI) (Morrison 2011b) (see §4.6).

2 Likelihood Ratio

The LR is the probability that the evidence would occur if an assertion is true, relative to the probability that the evidence would occur if the assertion is not true (Robertson & Vignaux 1995). Thus, the LR can be expressed as Equation 1).

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

For FVC, it will be the probability of observing the difference (referred to as the evidence, E) between the offender's and the suspect's speech samples if they had come from the same speaker (H_p) (i.e. if the prosecution hypothesis is true) relative to the probability of observing the same evidence (E) if they had been produced by different speakers (H_d) (i.e. if the defence hypothesis is true). 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).

For example, an LR of 20 means that the evidence (= the difference between the offender and suspect speech samples) is 20 times more likely to occur if the offender and the suspect had been the same individual than if they had been different individuals. Note that an LR value of 20 does NOT mean that the offender and the suspect are 20 times more likely to be the same person than different people, given the evidence.

The important point is that the LR is concerned with the probability of the evidence, given the hypothesis (either prosecution or defence), which is the province of forensic scientists, while the trier-of-fact is concerned with the probability of the hypothesis (either prosecution or defence), given the evidence. That is, the ultimate decision as to whether the suspect is guilty or not (e.g. the offender and suspect samples are from the same speaker 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) in order to assist the trier-of-fact to make a final decision (Morrison 2009: 229).

3 Database, target segment, and speakers

In this study, we used the monologues from the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al. 2000). There are two types of monologues in CSJ: Academic Presentation Speech (APS) and Simulated Public Speech (SPS). Both types were used in this study. APS was recorded live at academic presentations, most of them 12-25 minutes long. SPS contains 10-12 minute mock speeches on everyday topics.

For this study, we focused on the filler /e:/ and the /e:/ segment of the filler /e:to:/. Fillers are a sound or a word (e.g. *um*, *you know*, *like* in English) which is uttered by a speaker to signal that he/she is thinking or hesitating. We decided to use these fillers because 1) they are two of the most frequently used fillers (thus many monologues contain at least ten of these fillers) (Ishihara 2010), 2) the vowel /e/ reportedly has the strongest speaker-discriminatory power out of the five Japanese vowels /a, i. u, e, o/ (Kinoshita 2001), and 3) 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 by the speaker and carry no lexical meaning. They are thus not likely to be affected by the

pragmatic focus of the utterance. This is another reason we decided to focus on fillers in this study.

For the experiments, we selected our speakers based on five criteria: 1) availability of two non-contemporaneous recordings per speaker, 2) high spontaneity of the speech (e.g. not reading), 3) speaking entirely in standard modern Japanese, 4) containing at least ten /e:/ segments, and 5) availability of complete annotation of the data. Having real casework in mind, we selected only male speakers. This is because they are more likely to commit a crime than females (Kanazawa & Still 2000). These criteria resulted in 236 recordings (118 speakers x 2 non-contemporaneous recordings), and they were used in our experiments.

These 118 speakers (D_{all}) were divided into three mutually-exclusive sub databases; test database (D_{test} = 40 speakers), the background database ($D_{background}$ = 39 speakers) and the development database ($D_{development}$ = 39 speakers). Each speaker of these databases has two recordings which 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). The statistics which are necessary for conducting Monte Carlo simulations were calculated from these databases.

The test database was used to assess the performance of the FVC system. The background database was for a background population, and the development database was for obtaining the logistic-regression weight, which was used to calibrate the LRs of the test database (refer to §4.5 for the 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 which is used in many voice-related applications, including automatic speaker recognition. All original speech samples were downsampled to 16KHz, and then MFCC values were extracted from the mid-duration-point of the target segment /e:/ with a 20 ms wide hamming window. No normalisation procedure (e.g. Cepstrum Mean Normalisation) was employed as all recordings were made using the same equipment in CSJ.

4.2 General experimental design

There are two types of tests for FVC: one is the so-called *Same Speaker Comparison* (SS comparison) where two speech samples produced by the same speaker are expected to receive the desired LR value given the same-origin, whereas the other is, *mutatis mutandis*, *Different Speaker Comparison* (DS comparison).

For example, from the 40 speakers of the test database (D_{test}), 40 SS comparisons and 1560 independent (e.g. not-overlapping) DS comparisons are possible.

4.3 Likelihood ratio calculation

The LR of each comparison was estimated using the Multivariate Likelihood Ratio (MVLR) formula, which is one of the standard formulae used in FVC (Ishihara & Kinoshita 2008, Rose 2006, Morrison & Kinoshita 2008, Rose et al. 2004). Although the reader needs to refer to Aitken and Lucy (2004) for the full mathematical exposition of the MVLR formula, this formula estimates a single LR from multiple variables (e.g. 16 MFCC), discounting the correlation among them.

The numerator of the MVLR formula calculates the likelihood (= probability) of evidence, which is the difference between the offender and suspect speech samples, when it is assumed that both of the samples have the same origin (or the prosecution hypothesis (H_p) is true). For that, you need the feature vectors of the offender and suspect samples and the within-group (= speaker) variance, which is given in the form of a variance/covariance matrix. The same feature vectors of the offender and suspect samples and the between-group (= speaker) variance are used in the denominator of the formula to estimate the likelihood of getting the same evidence when it is assumed that they have different origins (or the defence hypothesis (H_d) is true). These within-group and between-group variances are estimated from the background dataset ($D_{background}$). The MVLR formula assumes normality for within-group variance while it uses a kernel-density model for between-group variance.

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. Thus, we can use a maximum of ten feature vectors to model each session of each speaker. In this study, we randomly generated X feature vectors (X = {2,4,6,8,10}) for each ses-

sion of each speaker 300 times using the normal distribution function modelled with the mean vector (μ) and variance/covariance matrix (ε) obtained from the original databases ({D_{test}, $D_{background}$, $D_{development}$}).

Figure 1 is an example showing 300 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.

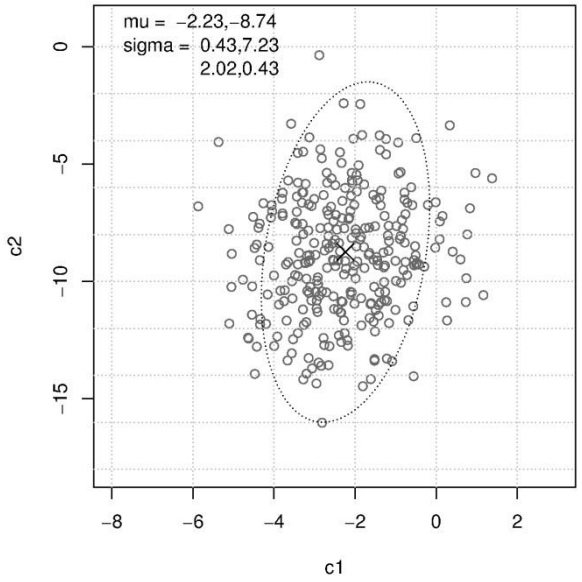

Figure 1: 300 randomly generated values (c1 and c2) from the statistics (μ and ε) obtained from the first session of the first speaker of the test database (only the first and second MFCC) and an ellipse. The cross = μ.

Experiments were repeatedly conducted using randomly generated feature vectors, as explained above. Two experiments: Experiments 1 and 2 were conducted in this study. In Experiment 1, we investigated how the token number (the number of feature vectors) of each speaker's session affects the performance of the FVC system. In Experiment 1, the same token number ({2,4,6,8,10}) was used across the test, background and development databases.

In Experiment 2, Experiment 1 was repeated with different token numbers in the background database ({2,4,6,8,10}) with the token number of the test and development databases kept constant. The aim of Experiment 2 was to investigate how the number of tokens in the background database affects the performance of the FVC system.

4.5 Calibration

A logistic-regression calibration (Brümmer & du Preez 2006) was applied to the derived LRs from the MVLR formula. Given two sets of LRs derived from the SS and DS comparisons and a decision boundary, calibration is a normalisation procedure involving linear monotonic shifting and scaling of the LRs relative to the decision boundary so as to minimise a cost function. 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 FVC system was assessed in terms of its validity (= accuracy) and reliability (= precision) using the log-likelihood-ratio cost (C_{llr}) and the 95% credible intervals (CI) as the metrics of validity and reliability, respectively.

Suppose that you have speech samples collected from two speakers at two different sessions which are denoted as S1.1, S1.2, S2.1, and S2.2, where S = speaker, and 1 & 2 = the first and second sessions (S1.1 refers to the first session recording collected from (S)peaker1, and S1.2 the second session from that same speaker). From these speech samples, two independent (not overlapping) DS comparisons are possible; S1.1 vs. S2.1 and S1.2 vs. S2.2. Further suppose that you conducted two separate FVC tests in the same way, but using two different features (Features 1 and 2), and that you obtained the $\log_{10}$LRs given in Table 1 for these two DS comparisons.

DS comparison	Feature 1	Feature 2
S1.1 vs. S2.1	-3.5	-2.1
S1.2 vs. S2.2	-3.3	0.2

Table 1: Example LRs used to explain the concept of validity and reliability.

Since the comparisons given in Table 1 are DS comparisons, the desired $\log_{10}$LR value would 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 Feature 1, both of the comparisons received $\log_{10}LR < 0$ while for Feature 2, only one of them got $\log_{10}LR < 0$. Feature 1 is better not only in that both $\log_{10}$LR values are smaller than 0

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

(supporting the correct hypothesis) but also in that they are further away from unity ($\log_{10}$LR = 0) than the $\log_{10}$LR values of Feature 2. Thus, it can be said that the validity (= accuracy) of Feature 1 is higher than that of Feature 2. This is the basic concept of validity.

Morrison (2011b: 93) argues that classification-accuracy/classification-error rates, such as equal error rate (EER), are inappropriate for use within the LR framework because they implicitly refer to posterior probabilities – which is the province of the trier-of-fact – rather than LRs – which is the province of forensic scientists – and "they are based on a categorical threshholding, error versus non-error, rather than a gradient strength of evidence." In this study, the log-likelihood-ratio cost (C_{llr}), which is a gradient metric based on LR for assessing the validity of the system performance was used. See Equation 2) for calculating C_{llr} (Brümmer & du Preez 2006). In Equation 2), N_{Hp} and N_{Hd} are the numbers of SS and of DS comparisons, and LR_i and LR_j are the LRs derived from the SS and DS comparisons, respectively. If the system is producing desired LRs, all the SS comparisons should produce LRs greater than 1, and the DS comparisons should produce LRs less than 1. In this approach, LRs which support counter-factual hypotheses are given a penalty. The size of this penalty is determined according to how significantly the LRs deviate from the neutral point. That is, an LR supporting a counter-factual hypothesis with greater strength will be penalised more heavily than the ones which are closer to unity, because they are more misleading. The FoCal toolkit[1] was also used for calculating C_{llr} in this study (Brümmer & du Preez 2006). The lower the C_{llr} value is, the better the performance.

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

Both of the DS comparisons given in Table 1 are the comparisons between S1 and S2. Thus, you can expect that the LR values obtained for these two DS comparisons should be similar as they are comparing the same speakers. However, you can see that the $\log_{10}$LR values based on Feature 1 are closer to each other (-3.5 and -3.3) than those based on Feature 2 (-2.1 and 0.2). In other words, the reliability (= precision) of Feature 1 is higher than that of Feature 2. This is the basic concept of reliability. As a metric of reliability, we used credible intervals, the Bayesian analogue of frequentist confidence intervals (Morrison 2011b). In this study, we calculated 95% credible intervals (CI) in the parametric manner based on the deviation-from-mean values collected from all of the DS comparison pairs. For example, CI = 1.23 and $\log_{10}$LR = 2 means that it is 95% certain that it is at least $\log_{10}$LR =

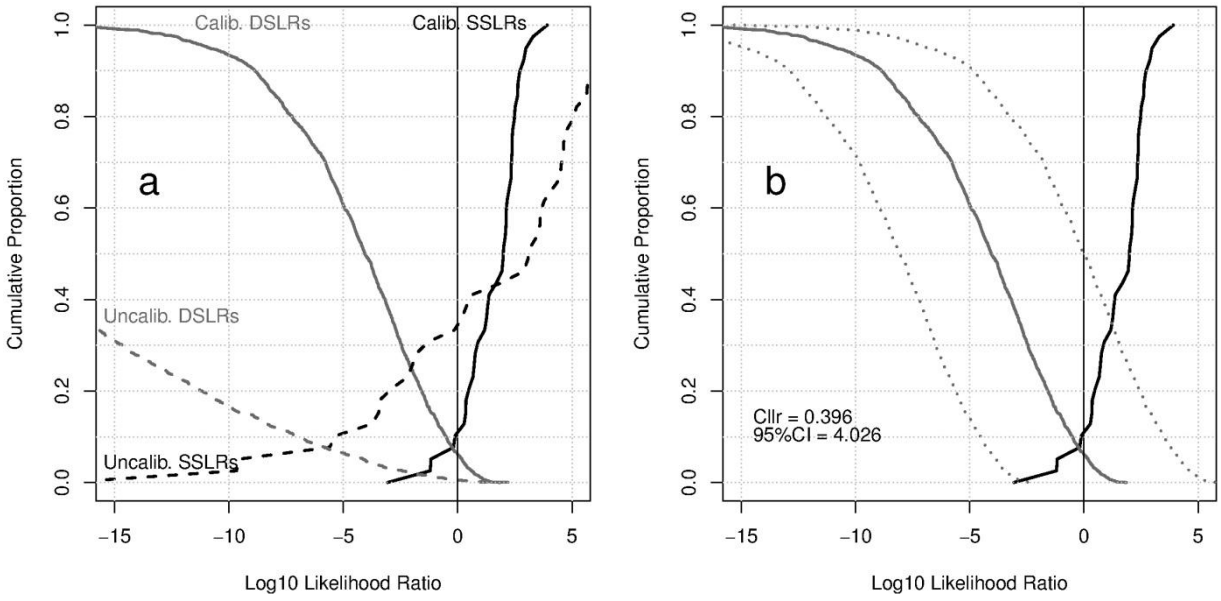

Figure 2: Tippett plot showing the uncalibrated (dashed curves) and calibrated (solid curves) LRs plotted separately for the SS (black) and DS (grey) comparisons (a), and Tippett plot showing the calibrated LRs with ±95% CI band (grey dotted lines) superimposed on the DS LRs (b). X-axis = $\log_{10}$LR; Y=axis = cumulative proportion. C_{llr} value was calculated from the calibrated LRs and CI value was calculated only for the calibrated DS LRs.

0.77 (= 2-1.23) and it is not greater than $\log_{10}$LR = 3.23 (= 2+1.23) for this particular comparison. The smaller the credible intervals, the better the reliability is.

Before presenting the results of Experiments 1 and 2, we conducted an experiment using the original databases (D_{test}, $D_{background}$, $D_{development}$). The results of this experiment are given as Tippett plots in Figure 2 with the C_{llr} and CI values. In these Tippett plots, the $\log_{10}$LRs, which are equal to or greater than the value indicated on the X-axis, are cumulatively plotted, separately for the SS and DS comparisons. 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 2a, calibrated and uncalibrated LRs are plotted together in order to show what sorts of effect the logistic-regression calibration brings to the uncalibrated LRs, and in Figure 2b, the calibrated LRs are plotted together with ±CI band on the DS LRs.

Theoretically speaking, the crossing point of the SS and DS LRs should be on $\log_{10}$LR = 0, but you can see the crossing point of the uncalibrated SS and DS LRs are far away from it in Figure 2b. In this circumstance, it is difficult to interpret the given LR appropriately as the theoretical threshold ($\log_{10}$LR = 0) and the obtained threshold ($\log_{10}$LR = ca. -7 in the uncalibrated LRs of Figure 2b) are completely different. A calibration technique needs to be applied in this situation. Please note that the calibrated SS and DS LRs given in Figure 2 are very well calibrated. The C_{llr} value was calculated using these calibrated SS and DS LRs, and it was 0.396. The CI was calculated based on calibrated DS LRs, and it was 4.026.

5 Experimental Results and Discussions

The results of Experiment 1 are graphically presented in Figure 3 in terms of C_{llr} and CI. In Figure 3a, the C_{llr} and CI values obtained from the Monte Carlo simulations (repeated 300 times) are plotted altogether with their mean values for each of the five different token numbers ({2,4,6,8,10}). The numerical values for the mean values are given in Table 2 together with their standard deviation (sd) values. Please note that the same token number was used across the test, background and development databases (test = background = development = {2,4,6,8,10}) in Experiment 1.

What we can observe from Figure 3a and Table 2 is that the validity of the system (C_{llr}) improves as the token number increases whereas the reliability of the system (CI) deteriorates. That is, there is a trade-off between the validity and reliability of the system. The improvement in validity as a function of the token number is non-linear in that there is a large improvement from the token number = {2} to {4} (0.66->0.51)

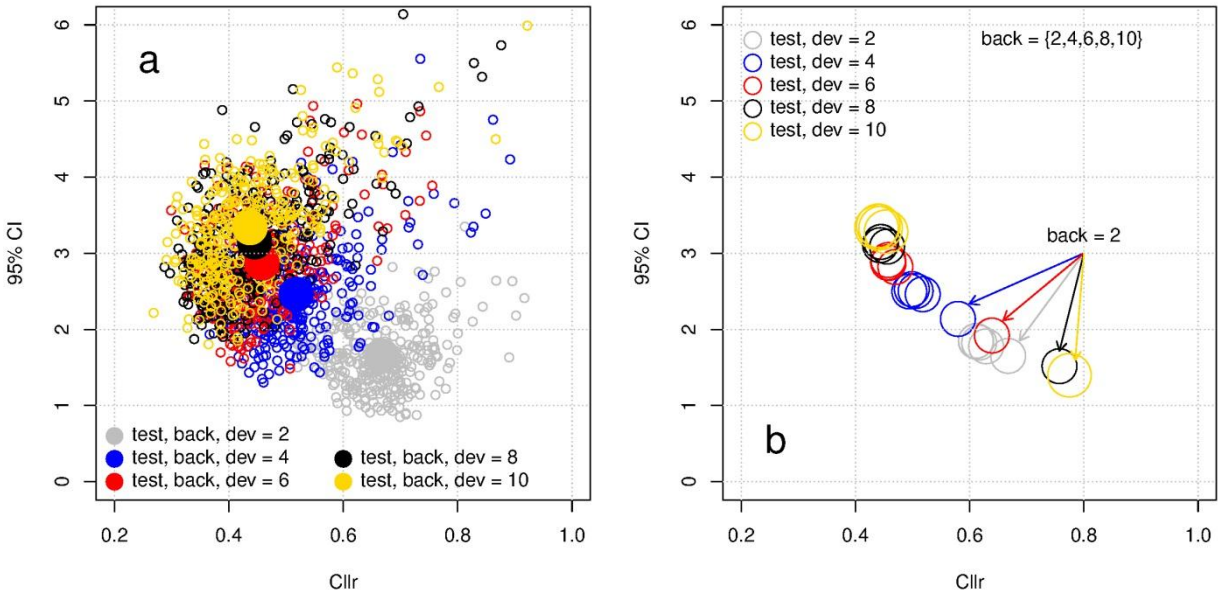

Figure 3: The C_{llr} and CI values of the 300 repeated Monte Carlo simulations are plotted separately for the different token numbers {2,4,6,8,10} with their mean values (large filled circles) (a). The mean C_{llr} and CI values of the 300 repeated Monte Carlo simulations (big empty circles) differing in the token numbers ({2,4,6,8,10}) of the background database (b). X-axis = C_{llr}; Y-axis = CI; test, back and dev = test, background and development databases.

whereas there is not much improvement between the token number = {6} and the token number = {10} (0.45->0.44->0.43). That is, if you have six repeated tokens (e.g. six *yes* tokens for each session of each speaker) in the databases, the performance of the system can be expected to be as good as when you have as many as ten repeated tokens.

	test = background = development =				
	2	4	6	8	10
C_{llr}	0.66	0.51	0.45	0.44	0.43
sd	0.073	0.087	0.091	0.093	0.090
CI	1.65	2.46	2.87	3.14	3.33
sd	0.427	0.629	0.711	0.734	0.700

Table 2: The numerical values of Figure 3a (only mean values).

Another observation that can be made is that the C_{llr} and CI values are more widely scattered when the token number is {6,8,10} than {2,4}. This point can be seen in the sd values given in Table 2 in that, for example, the sd values of the C_{llr} and CI are far smaller when the token number is {2} (0.073 and 0.427, respectively) than when the token number is {10} (0.090 and 0.700, respectively). That is, the performance of the system widely fluctuates when the token number is high (e.g. {6,8,10}).

In Experiment 2, Experiment 1 was repeated five times with the five different token numbers ({2,4,6,8,10}) in the background database. The results of Experiment 2 are given in Figure 3b in which only the mean C_{llr} and CI values are plotted in order to prevent the figure from becoming too crowded. The numerical values of Figure 3b are given in Table 3. For example, the experiment with the token number of {10} in the test and development databases was repeated five times, differing the token number in the background database (background = {2,4,6,8,10}), and then the mean C_{llr} and CI values of these five experiments are plotted in the same colour (gold for the token number of {10} in the test and development databases) in Figure 3b.

We can observe from Figure 3b and Table 3 that each experimental set (e.g. test = development = 8, background = {2,4,6,8,10}) has one result which is very different in performance from the other four results. For example, the results of the token number of {10} in the test and development databases with the token numbers of {4,6,8,10} in the background database are more or less the same (C_{llr} = ca. 0.44 and CI = ca. 3.3) whereas they are significantly better in terms of C_{llr} than the result with the token number of {2} in the background database (= 0.77). In fact, regardless of the token number in the test and development databases, the performance of the system is worse when there are only two repeated tokens in the background database than when there are four or more repeated tokens ({4,6,8,10}) (refer to the arrows given in Figure 3b).

test = dev =	back =	C_{llr}	CI
2	2	0.66	1.65
	4	0.62	1.77
	6	0.61	1.82
	8	0.61	1.84
	10	0.61	1.84
4	2	0.57	2.13
	4	0.51	2.46
	6	0.50	2.50
	8	0.49	2.52
	10	0.49	2.49
6	2	0.63	1.91
	4	0.46	2.82
	6	0.45	2.87
	8	0.45	2.88
	10	0.45	2.91
8	2	0.75	1.51
	4	0.45	3.08
	6	0.44	3.10
	8	0.44	3.14
	10	0.44	3.14
10	2	0.77	1.39
	4	0.45	3.28
	6	0.44	3.33
	8	0.43	3.36
	10	0.43	3.33

Table 3: The numerical values of Figure 3b.

Furthermore, this difference in performance between the token numbers of {4,6,8,10} and that of {2} in the background database becomes greater as the number of tokens used in the test and development databases increases. For example, as can be seen in Table 3, the difference in question is relatively small for the test and development databases = {2} (C_{llr} = 0.66 and CI = 1.65 for the background = {2}; average C_{llr} = 0.61 and average CI = 1.81 for the background = {4,6,8,10}) whereas it is far larger for the test and development databases = {10} (C_{llr} = 0.77 and CI = 1.39 for the background = {2}; average C_{llr} = 0.43 and average CI = 3.32 for the background = {4,6,8,10}).

As far as the C_{llr} values are concerned, the performance never deteriorates as the size increases from the background = {4} to {10}. Whereas there are some very small fluctuations in performance in terms of the CI values from the background = {4} to {10}. The reasons for these fluctuations are not clear at this stage.

The results of Experiment 2 tell us that, if you have four repeated tokens (e.g. four *yes* tokens for each session of each speaker) in the background database, the system can achieve as good a performance as when you have ten repeated tokens. However, if you have only two repeated tokens in the background database, it will result in an underperformance of the system in comparison to when you have four or more repeated tokens.

6 Conclusions and Future Directions

This study investigated how the offender and suspect sample sizes (or the within-speaker sample size) influences the performance of an FVC system. In order to answer this question, two experiments based on Monte Carlo simulations: Experiments 1 and 2, were conducted.

In Experiment 1, five different token numbers ({2,4,6,8,10}) were used in the databases to see how the performance of the system would be influenced by the token number. The results demonstrated that 1) there was a trade-off between the validity (C_{llr}) and reliability (CI) of the system; 2) there was a large improvement in the validity between the token number = {2} and the token number = {4} whereas no large improvement was observed from the token number = {6} to the token number = {10}. That is, if we have six repetitions of the target segment/word (e.g. *yes*), the system validity is almost as good as when we have ten repetitions.

In Experiment 2, Experiment 1 was repeated by changing the token number ({2,4,6,8,10}) of the background database while keeping the same token number for the test and development databases. The results of Experiment 2 demonstrated that regardless of the token number in the test and development databases, the system with the token number = {2} in the background database significantly underperformed in accuracy when compared to the systems with the token number = {4,6,8,10}, of which the performances were very similar. The results of Experiment 2 also demonstrated that the above-mentioned discrepancy in performance between two repeated tokens ({2}) and four or more repeated tokens ({4,6,8,10}) becomes wider as the token number of the test and development databases increases.

These results suggest that when we compile a database which can be used as background population data, we do not need many repetitions in the database as a model based on four repeated tokens can achieve very similar results as one based on ten repeated tokens. However, if we have only two repeated tokens in the background database, we need to be aware that the performance will be compromised, even if you have many repetitions in the test and development databases.

In this study, we mainly focused on the token numbers of the test and background databases. However, it goes without saying that the token number of the development database is also important to the performance of a system. We need to look into this point as well.

In this study, although some other techniques are available for the estimate of LRs, the MVLR formula was used. For example, Morrison (2011a) reported that the procedures based on the Gaussian Mixture Model – Universal Background Model (GMM-UBM) outperformed those based on MVLR procedures, and that the GMM-UBM resulted in an improvement in both the validity and reliability (without trade-offs between them). Since the GMM-UBM is another popular way of estimating LRs in FVC, it is important to investigate the relationship between its performance and the sample size as well.

Acknowledgments

The author appreciates the very detailed comments and suggestions made by the three anonymous reviewers.

References

Aitken CGG & D Lucy 2004 'Evaluation of trace evidence in the form of multivariate data' *Journal of the Royal Statistical Society Series C-Applied Statistics* 53: 109-122.

Aitken CGG & DA Stoney 1991 *The Use of Statistics in Forensic Science* Ellis Horwood New York; London.

Aitken CGG & F Taroni 2004 *Statistics and the Evaluation of Evidence for Forensic Scientists* Wiley Chichester.

Bozza S, F Taroni, R Marquis & M Schmittbuhl 2008 'Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship' *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57(3): 329-341.

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

Fishman GS 1995 *Monte Carlo: Concepts, Algorithms, and Applications* Springer New York.

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 & Y Kinoshita 2008 'How many do we need? Exploration of the population size effect on the performance of forensic speaker classification' *Proceedings of Interspeech 2008*: 1941-1944.

Kanazawa S & MC Still 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 Ph.D. thesis, the Australian National University.

Kinoshita Y, S Ishihara & P Rose 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.

Kinoshita Y & M Norris 2010 'Simulating spontaneous speech: Application to forensic voice comparison' *Proceedings of the 13th Australasian International conference on Speech Science and Technology*: 26-29.

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

Morrison GS 2009 'Forensic voice comparison and the paradigm shift' *Science & Justice* 49(4): 298-308.

Morrison GS 2011a '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* 53(2): 242-256.

Morrison GS 2011b 'Measuring the validity and reliability of forensic likelihood-ratio systems' *Science & Justice* 51(3): 91-98.

Morrison GS & Y Kinoshita 2008 'Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English vertical bar o vertical bar Formant Trajectories' *Proceedings of Interspeech 2008*: 1501-1504.

Neumann C, C Champod, R Puch-Solis, N Egli, A Anthonioz & A Bromage-Griffiths 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 & GA Vignaux 1995 *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* Wiley Chichester.

Rose P 2006 'Technical forensic speaker recognition: Evaluation, types and testing of evidence' *Computer Speech and Language* 20(2-3): 159-191.

Rose P, D Lucy & T Osanai 2004 'Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects model: A "non-idiot's Bayes" approach' *Proceedings of the 10th Australian International Conference on Speech Science and Technology*: 492-497.