Replicate Mismatch between Test and Background/Development Databases: The Effect on the Performance of Likelihood Ratio-based Forensic Voice Comparison

Shunichi Ishihara

Department of Linguistics, the Australian National University

shunichi.ishihara@anu.edu.au

Abstract

This study reports the extent that mismatch in within-speaker replicate numbers between test and background/development databases causes an influence on the performance of a forensic voice comparison (FVC) system. FVC tests are repeatedly carried out using the Monte Carlo simulation technique with temporal MFCC features and the Multivariate Kernel Density Likelihood Ratio Procedure. The performance of the FVC system is assessed in terms of its validity (C_{llr}) and reliability (*CI*) under various mismatch conditions. It is shown in this study that a matched replicate number across the three databases yields the best result in terms of C_{llr}. However, the performance of *CI* improves as the replicate number of the test database becomes smaller.

Index Terms: likelihood ratio, forensic voice comparison, database mismatch, replicate number, Monte Carlo simulation

1. Introduction

Data mismatch is a common problem in forensic voice comparison (FVC) casework. This problem includes not only mismatch between offender and suspect samples, but also mismatch between the offender/suspect samples and the samples of the background/development databases, typically in terms of speaking style, transmission channel, sample size, etc. Although it is virtually inevitable to have some data mismatches in real casework, it is widely acknowledged that mismatches should be avoided as much as possible to achieve optimal results, and it also has been empirically demonstrated that database mismatch has a critical effect on system performance [1, 2].

In FVC experiments, three different databases: namely test, background and development databases, are usually used. A test database, which imitates the offender-suspect comparisons of real casework, is used to assess the performance of an FVC system. A background database is used to estimate a model of the distribution of measured acoustic properties in the relevant population. A development database is typically used to calculate weights for logistic-regression calibration. In this study, we focus on the mismatch between test and background/development databases in terms of the within-speaker sample size, or more specifically, the within-speaker replicate number.

Suppose that an FTC caseworker is working on casework for which he/she has carefully listened to the provided offender and suspect samples, and judged that their speaking style, transmission channel, and so on, are comparable. Furthermore, the caseworker has identified six *yes* tokens each from the suspect and offender samples, and assessed that they can be compared for analysis. The caseworker has also managed to compile appropriate background and development databases. However, he/she only has two tokens of *yes* for each session of the speakers included in the background and development databases – a mismatch with the test database in terms of the within-speaker replicate number. This study takes such a scenario into consideration.

The current study investigates the extent to which the performance of an FVC system is compromised under mismatched conditions between the test and background/development databases. In order to achieve this, the performance of an FVC system, based on temporal MFCC features and the Multivariate Kernel Density Likelihood Ratio (MVKD) procedure, is tested under the above-mentioned mismatched conditions. The system performance is repeatedly tested using the Monte Carlo technique, and assessed by means of validity [3] and reliability [4].

2. Likelihood ratio

The current study is a likelihood ratio (LR) based FVC study. For FVC, as expressed in Equation 1). LR is 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) [5].

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

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.

3. Database, target segment, and speakers

In this study, we used monologues from the Corpus of Spontaneous Japanese (CSJ) [6]. 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:/. 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) [7], 2) the vowel /e/ reportedly has the strongest speaker-discriminatory power out of the five Japanese vowels /a, i. u, e, o/ [8], 3) the segment /e:/ is significantly long, so it is easy

to extract stable spectral features, and 4) it is believed that fillers are uttered unconsciously by the speaker and carry no lexical meaning. They are therefore not likely to be affected by the pragmatic focus of the utterance.

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. We selected only male speakers. This is because they are more likely to commit a crime than females [9]. These criteria resulted in 236 recordings (118 speakers x 2 non-contemporaneous recordings) for use in our experiments.

These 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 non-contemporaneous recordings. 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 necessary for conducting Monte Carlo simulations (mean vector μ and variance/covariance matrix ε) were calculated from these databases.

There are two types of tests for FVC: one type is *Same Speaker Comparisons* (SS comparisons), in which two speech samples produced by the same speaker are expected to receive the desired LR value (LR > 1) given the same-origin, and the other type is, *mutatis mutandis*, *Different Speaker Comparisons* (DS comparisons). From the 40 speakers of the test database, 40 SS comparisons and 1560 independent (e.g. non-overlapping) DS comparisons are possible.

4. Experiments

4.1. Features

We used 16 Mel Frequency Cepstrum Coefficients (MFCC) as feature vectors in the experiments. 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.

4.2. Likelihood ratio calculations

The LR of each comparison was estimated using the Multivariate Kernel Density Likelihood Ratio (MVKD) procedure, which is one of the standard formulae used in FVC [10-13]. The MKVD formula estimates a single LR from multiple variables (e.g. 16 MFCCs), discounting the correlation among them (refer to [14] for a full mathematical exposition of the formula).

4.3. Repeated experiments using Monte Carlo simulations

In order to investigate the effect of replicate number mismatch between test and background/development databases, we conducted a series of experiments with different replicate numbers ({2,4,6,8,10}) in the test database, while keeping the replicate number of the background/development databases constant (either {2},{4},{6},{8}, or {10}). Table 1 contains all experimental combinations carried out in the current study, and they are given according to experimental sets. For example, for the experimental set 1), which had a constant replicate number of {2} in the background/development databases, we conducted five different experiments by changing the replicate number of the test database (either {2},{4},{6},{8} or {10}). Thus, 25 different experiments (= 5 experiments * 5 experimental sets) were carried out overall.

Table 1. *All possible experimental combinations. ex. = experimental sets; test = the replicate numbers of test database; back./dev. = The replicate number of background/development databases.*

ex.	test		back./dev.
set 1)	{2},{4},{6},{8} and {10}	vs.	{2}
set 2)	{2},{4},{6},{8} and {10}	vs.	{4}
set 3)	{2},{4},{6},{8} and {10}	vs.	{6}
set 4)	{2},{4},{6},{8} and {10}	vs.	{8}
set 5)	{2},{4},{6},{8} and {10}	vs.	{10}

As explained earlier, each speaker has two sets of ten /e:/ segments, and 16 MFCC values were extracted from each /e:/ segment. 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 session 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 test, background and development databases. Thus, each of the 25 experiments listed in Table 1 was repeated 300 times using the Monte Carlo technique.

4.4. Calibration

Theoretically speaking, the LRs estimated by the MVKD formula should be well-calibrated. However, this is not always the case when the modelling assumptions of the formula are violated. For example, although within-speaker variance is assumed to be constant in this formula, this is obviously not an appropriate assumption for speech acoustics. Thus, the poorly-calibrated LRs estimated by the MVKD formula, which are customarily referred to as *scores*, need to be calibrated.

Logistic-regression calibration [3] was applied to the outcomes of the MVKD formula in the current study. Given two sets of LRs (or *scores*) derived from 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 [3]. The logistic-regression weight was obtained from the development database, as explained earlier.

4.5. 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}) [3] and 95% credible intervals (*CI*) [4], respectively. We calculated the *CI* using the non-parametric method on the DS comparison pairs.

5. Experiment results and discussion

The experimental results are graphically presented in Figure 1 in terms of C_{llr} and *CI*. In Figure 1, the mean C_{llr} and *CI* values obtained from the Monte Carlo simulations (repeated 300

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

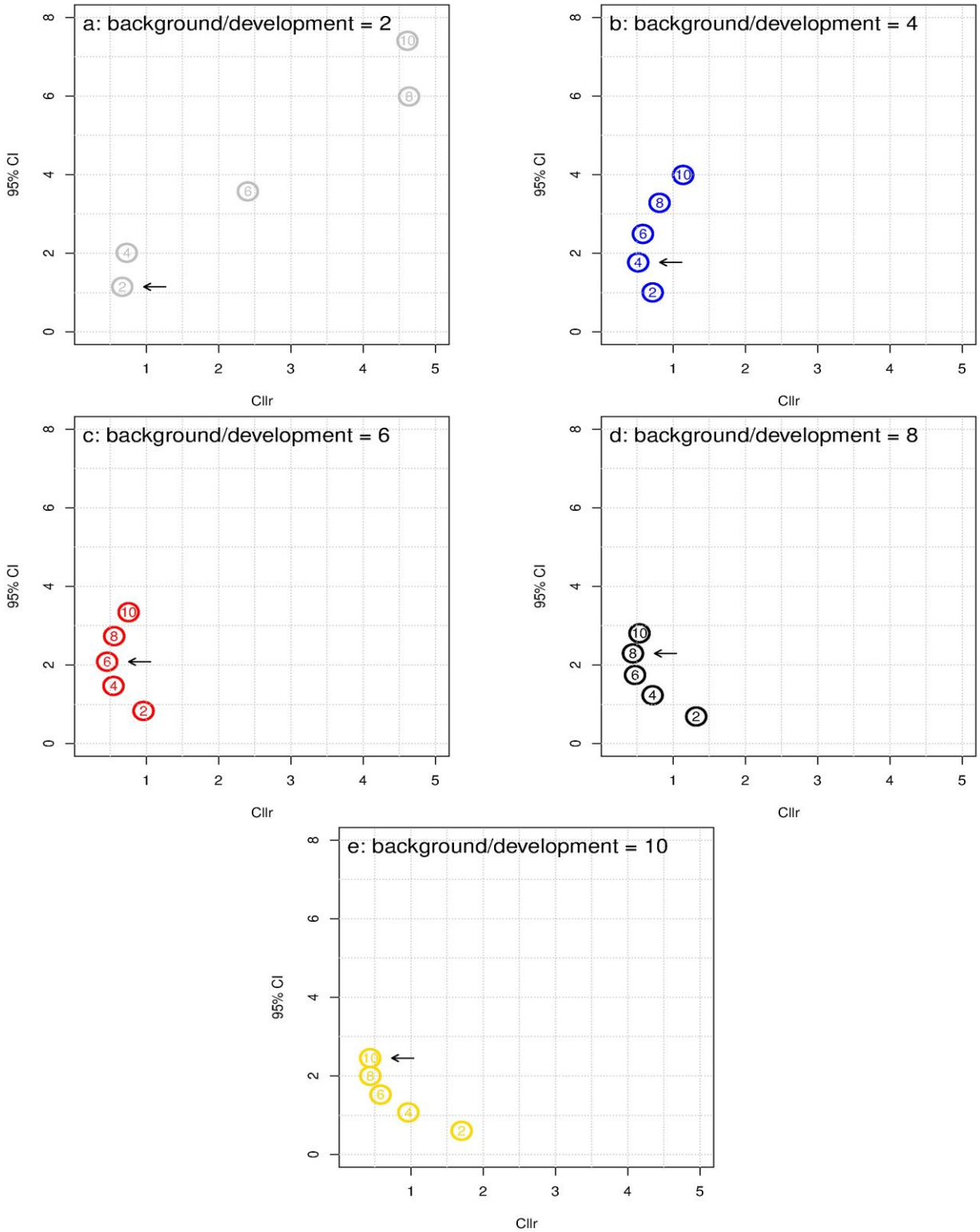

Figure 1: *The results of the five different experimental sets given in Table 1 in terms of mean C_{llr} (x-axis) and CI (y-axis). The replicate number of background/development databases = 2 (a), 4 (b); 6 (c); 8 (d) and 10 (e). The number appearing in a circle = the replicate number of the test database. Arrows = the results of the matched databases.*

times for each experiment) are plotted for each experiment set given in Table 1, but separately for the replicate number ({2,4,6,8,10}) of the test database. For example, Figure 1a shows the mean C_{llr} and *CI* values of the experimental set 1), in which each session of each speaker has a constant replicate number of two in the background/development databases, but five different replicate numbers for the test database. The numbers in circles indicate the replicate numbers of the test database.

We can observe from Figure 1 that when the replicate number is identical across the three databases (in other words, the matched experiments), the best performance in terms of C_{llr} can be achieved within each experimental set, regardless of the number of replicates. The matched experiments are indicated by arrows in Figure 1. In terms of *CI*, the system performs best when the replicate number of the test database = {2}, regardless of the replicate number of background/development databases, and the *CI* value increases as the replicate number increases.

The magnitude of the difference in replicate number between the test and background/development databases brings different influences on the performance of the system. Although it is somewhat expected, the larger the degree of mismatch, the greater the influence on the performance of the system. For example, in Figure 1a, in which the results of the experimental set 1) are given, the mean C_{llr} and *CI* values of the test database = {6} (the degree of mismatch is 4) is located further away from those of the test database = {2} (the matched case), than those of the test database = {4} (the degree of mismatch is 2). Furthermore, the effect of replicate mismatch becomes weaker as the replicate number of the background/development databases increases. Most notably, when the replicate number of the background/development databases = {2}, the mismatch will bring a large influence on the performance of the system. This point can be clearly seen by comparing Figure 1a, which are the results of the experimental set 1), and Figure 1bcde, which are the results of the experimental sets 2), 3), 4) and 5). The circles in Figure 1a are further apart from each other than in Figure 1bcde. One of the reviewers pointed out that the very high C_{llr} values occurring for some of the mismatched conditions are possibly due to the computational issues with the MVKD [15]. Nevertheless, the results of the current study are very consistent.

Another observation that can be made from Figure 1bcd is that in cases where there is a difference of 2 in replicate number between the test and background/development databases, there is no clear difference in C_{llr}, but there is a clear difference in *CI* (the latter performs better than the former in terms of *CI*). For example, circle 4 of Figure 1c, in which the replicate number of the test database is smaller by 2 than that of the background/development databases, and circle 8 of Figure 1c, in which the replicate number of the test database is larger by 2 than that of the background/development databases, are similar in terms of C_{llr} (C_{llr}: circle 4 = 0.54; circle 8 = 0.55) but the former (*CI* = 1.46) has a lower *CI* value (thus better in performance) than the latter (*CI* = 2.73). That is, when there is a mismatch in replicate number between the test and background/development databases, and the mismatch is small (e.g. ±2 in replicate number), it is better that the test database has fewer replicate numbers than the background/development databases than the other way around.

6. Conclusions

This study showed the extent to which an FVC system is compromised under the conditions of replicate number mismatches between the test and background/development databases. We demonstrated that the system yielded the best results in terms of C_{llr} when the replicate number is matched across the three databases, regardless of the number of replicates. However, the current study also showed that this is not the case in terms of *CI* – the *CI* improves as the replicate number becomes smaller in the test database. This study also reported that when the replicate number of the background/development databases is small (e.g. {2}), the same magnitude of mismatch brings a larger influence on the performance than when the replicate number of the background/development databases is large (e.g. ≥ {4}).

7. Acknowledgements

The author greatly appreciates the very useful comments of the three anonymous reviewers.

8. References

[1] 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 & Language*, vol. 20, pp. 331-355, 2006.

[2] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with Ahumada III: A public real-casework database in Spanish," in *Proceedings of Interspeech 2008*, 2008, pp. 1493-1496.

[3] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230-275, Apr-Jul 2006.

[4] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, pp. 91-98, Sep 2011.

[5] B. Robertson and G. A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: Wiley, 1995.

[6] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of the 2nd International Conference of Language Resources and Evaluation*, 2000, pp. 947-952.

[7] S. Ishihara, "Variability and consistency in the idiosyncratic selection of fillers in Japanese monologues: Gender differences," in *Proceedings of the Australasian Language Technology Association Workshop 2010*, Melbourne, Australia, 2010, pp. 9-17.

[8] Y. Kinoshita, "Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants," PhD thesis, the Australian National University, 2001.

[9] S. Kanazawa and M. C. Still, "Why men commit crimes (and why they desist)," *Sociological Theory*, vol. 18, pp. 434-447, Nov 2000.

[10] S. Ishihara and Y. Kinoshita, "How many do we need? Exploration of the population size effect on the performance of forensic speaker classification," in *Proceedings of Interspeech 2008*, 2008, pp. 1941-1944.

[11] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech and Language*, vol. 20, pp. 159-191, Apr-Jul 2006.

[12] G. S. Morrison and Y. Kinoshita, "Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories," in *Proceedings of Interspeech 2008*, 2008, pp. 1501-1504.

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

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

[15] B. Nair, E. Alzqhoul, and B. J. Guillemin, "Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis," *International Journal of Speech Language and the Law*, vol. 21, pp. 83-112, 2014.