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

Shunichi Ishihara

Department of Linguistics, the Australian National University, Australia

shunichi.ishihara@anu.edu.au

Abstract

In this paper, we report on a study which demonstrates that the mismatch in within-speaker replicate numbers (the number of tokens used to model each sample) between test/background and development databases has a large impact on the performance of a forensic voice comparison (FVC) system. We describe how and to what extent the different degrees of the mismatch influence the performance of the FVC system. The performance of an FVC system based on temporal MFCC features and the Multivariate Kernel Density Likelihood Ratio procedure is tested in terms of its validity and reliability under the mismatched conditions. The Monte Carlo technique is employed to repeatedly carry out FVC tests. We report that the databases matched with respect to replicate numbers result in optimal performance in terms of validity, but not in terms of reliability.

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 between the offender/suspect samples and the background/development databases, typically in terms of speaking style, transmission channel, sample size, etc. Although it is almost inevitable to have some data mismatches in real casework, it is widely acknowledged that any 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 impact 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 consists of samples collected from the relevant population, and is used to build a distributional model of measured properties from the collected samples. A development database is typically used to calculate weights for logistic-regression calibration. In this study, we focus on the mismatch between test and development databases in terms of the within-speaker sample size, or more specifically, the within-speaker replicate number.

Suppose that you are working on casework for which you have carefully listened to the provided offender and suspect samples, and judged that their speaking style, transmission channel, and so on, are comparable. From the samples, you have identified six yes tokens from both the suspect and offender samples, and assessed that they can be compared for analysis. You have further managed to compile an appropriate background database conforming to the conditions of the offender and suspect samples. However, you only have two tokens of yes for each session of the speakers included in the development database – a mismatch with the test database in terms of the within-speaker replicate number.

The current study investigates the degree of impact that the mismatched conditions between the test and development databases will bring about in terms of the performance of an FVC system. For this purpose, the performance of an FVC system, based on temporal MFCC features and the Multivariate Kernel Density Likelihood Ratio (MVKD) procedure [3], 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 [4] and reliability [5].

As explained earlier, the typical use of a development database in FVC is to obtain weights for logistic-regression calibration. Clearly, in order to obtain the optimal calibration result, the development database should be comparable with the other databases. In other words, the current study investigates the extent to which the system performance is compromised in terms of a calibration result if the development database is mismatched to the other databases.

2. Likelihood ratio

The likelihood ratio (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 [6]. Thus, the LR can be expressed as Equation 1).

\[
LR = \frac{p(E|H_p)}{p(E|H_d)}
\]

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 (Hp) (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 (Hd) (i.e. if the defence hypothesis is true). The relative strength of the given evidence with respect to the competing hypotheses (Hp vs. Hd) 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) [7]. 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 /etə/.
Filblers 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) [8], 2) the vowel /e/ reportedly has the strongest speaker-discriminatory power out of the five Japanese vowels /a, i, u, e, o/ [9], 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. With real casework in mind, we selected only male speakers. This is because males are more likely to commit a crime than females [10]. 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 had 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, σ) were calculated from these databases.

The test database is used to assess the performance of the FVC system. The background database is used to estimate a model of the distribution of measured acoustic properties in the relevant population (refer to §4.2 for a more specific use of the background database in the MVKD procedure). The development database is used to obtain the logistic-regression weights, which were used to calibrate the LRs of the comparisons generated from the test database (refer to §4.5 for a detailed explanation of calibration).

There are two types of tests for FVC: one type is called 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 hammering window. No normalisation procedure (e.g. Cepstrum Mean Normalisation) was employed as all recordings were made using the same equipment in CSJ.

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 [11-14]. Although the reader needs to refer to [3] for a full mathematical exposition of the MVKD formula, this formula estimates a single LR from multiple variables (e.g. 16 MFCCs), discounting the correlation among them.

The numerator of the MVKD formula calculates the likelihood 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 (Hp) is true). For that, one needs the feature vectors of the offender and suspect samples and the within-speaker variance, given in the form of a variance/covariance matrix. 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 likelihood of getting the same evidence when it is assumed that they have different origins (or the defence hypothesis (Hd) is true). These within-speaker and between-speaker variances are estimated from the background dataset. The MVKD formula assumes normality for within-speaker variance while it uses a kernel-density model for between-speaker variance.

4.3. Repeated experiments using Monte Carlo simulations

In order to investigate the mismatch effect between test/background and development databases with regard to replicate numbers, we conducted a series of experiments with different replicate numbers (2,4,6,8,10) in the development database, while the replicate number of the test/development databases is kept constant (either 2,4,6,8,10). 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 constantly has a replicate number of 2 in the test/background databases, we conducted five different experiments by changing the replicate number of the development database (either 2,4,6,8 or 10). Thus, 25 different experiments (5 experiments x 5 experimental sets) were carried out altogether.

<table>
<thead>
<tr>
<th>ex.</th>
<th>dev.</th>
<th>test/back.</th>
</tr>
</thead>
<tbody>
<tr>
<td>set 1</td>
<td>2,4,6,8,10 vs. 2</td>
<td></td>
</tr>
<tr>
<td>set 2</td>
<td>2,4,6,8 vs. 4</td>
<td></td>
</tr>
<tr>
<td>set 3</td>
<td>2,4,6,8 vs. 6</td>
<td></td>
</tr>
<tr>
<td>set 4</td>
<td>2,4,6,8 vs. 8</td>
<td></td>
</tr>
<tr>
<td>set 5</td>
<td>2,4,6,8 vs. 10</td>
<td></td>
</tr>
</tbody>
</table>

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 (Σ) obtained from the original test, background and development. 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, it 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 [4] was applied to the outcomes of the MVKD formula in the current study. Note that many studies [13, 15, 16], in which the MVKD formula was used for the LR estimations, reported a substantial improvement in performance when the output of the MVKD formula was calibrated using logistic-regression calibration. Given two sets of LRs (or 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 LRs relative to the decision boundary so as to minimise a cost function. The FoCal toolkit was used for the logistic-regression calibration in this study [4]. 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_{LR}) [4] and 95% credible intervals (Cl) [5], respectively. We calculated the Cl using the non-parametric method on the DS comparison pairs. Note that the use of Cl as a performance metric is contentious. The author is aware that some people including one of the reviewers consider the use of Cl for an LR unsound. Nevertheless, the Cl is used as the metric of reliability in this study.

5. Experimental Results and Discussions

The experimental results are graphically presented in Figure 1 in terms of C_{LR} and Cl. In Figure 1, the mean C_{LR} and Cl values obtained from the Monte Carlo simulations (repeated 300 times for each experiment) are plotted for each experiment set given in Table 1, but separately for the replicate number (1,2,4,6,8,10) of the development database. For example, Figure 1a, shows the mean C_{LR} and Cl values of the experimental set 1, in which each session of each speaker constantly has two replicates for the test/background databases, but five different replicate numbers for the development database. The numbers in circles indicate the replicate numbers of the development database.

We can observe from Figure 1, first of all, that when the replicate number of the test/background databases is ≤6 (Figure 1a,b,c), the optimal result in terms of C_{LR} can be obtained when the matched databases were used. The different degrees of mismatch also bring different impacts to the performance. Although it is expected, the larger the degree of the mismatch is, the higher the deteriorative impact on the performance in terms of C_{LR}. However, it appears as though there are two different types of impact, depending on the nature of the mismatch. If the replicate number of the development database is smaller than that of the test/development databases (case 1), the result of the mismatch is a deterioration of the performance both in terms of C_{LR} and Cl. For example, the blue-circled 2 (Figure 1b), where test/background = 4 and development = 2, underperformed both in terms of C_{LR} (C_{LR} = 0.621; Cl = 2.211) in comparison to the blue-circled 4, which is matched in the replicate number across the three databases (C_{LR} = 0.517; Cl = 1.768).

If the replicate number of the development database is larger than that of the test/background databases (case 2), it still negatively influenced the performance in terms of C_{LR} (yet to a lesser degree compared to case 1), but it improved the performance in terms of Cl. For example, the blue-circled 6 (Figure 1b), where test/background = 4 and development = 6, underperformed in terms of C_{LR} (C_{LR} = 0.560) in comparison to the blue-circled 4 (C_{LR} = 0.517), but it outperformed (Cl = 1.443) the matched one (Cl = 1.768) in terms of Cl. The same observation can be made from Figure 1c.

When the replicate number of the test/development databases is ≥8, some interesting observations can be made. Figure 1d shows that the mismatch type of case 2 described above did not deteriorate the performance in terms of C_{LR} while it still resulted in an improvement of the performance in terms of Cl. The black-circled 10, where the replicate number of the development database is larger than that of the test/background databases by 2, performed just as well (C_{LR} = 0.440) in terms of C_{LR} as the black-circled 8 (C_{LR} = 0.444), which has matched databases, while it still made an improvement in terms of Cl (Cl = 2.293) in comparison to the black circle 8 (Cl = 2.293). This may indicate that when the replicate number is relatively large (≥8), the performance of the system can be improved in terms of Cl without compromising the performance in terms of C_{LR}.

Regardless of the replicate number of the test/development databases, the reliability (Cl) of the FVC system improved as the replicate number of the development database increased.

6. Conclusions

This study demonstrated that any mismatch in replicate number between test/background and development databases causes a significant impact to the performance of an FVC system in terms of both validity (C_{LR}) and reliability (Cl). We reported that when the replicate number is small (≤6), the matched databases result in the optimal performance in terms of validity (C_{LR}). However, when the replicate number is relatively large (≥8), it was pointed out that the use of a larger replicate number for the development database may improve the system reliability without compromising the validity. This point warrants further research.
7. Acknowledgements

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

8. References


