EFFECT OF SCORE SAMPLING ON SYSTEM STABILITY IN LIKELIHOOD RATIO BASED FORENSIC VOICE COMPARISON

Bruce Xiao Wang, Vincent Hughes and Paul Foulkes

Department of Language and Linguistic Science, University of York, U.K.
{xw961/vincent.hughes/paul.foulkes}@york.ac.uk

ABSTRACT

In forensic voice comparison (FVC) cases it is essential to make sure that conclusions are reliable, robust, and replicable. This is especially true for data-driven FVC that relies on databases of speakers to estimate empirically the strength of the voice evidence. A key issue for such approaches is the validity of likelihood ratio (LR) output according to the specific speakers used for training and testing systems. The present study addresses this issue using simulated scores with different score distributions for training and test data. Experiments were replicated 100 times by varying the sampling of (1) both training and test scores, (2) training scores only, and (3) test scores only. The results show that sampling both test and training scores yielded the largest system variability with C_{llr} varying from 0.03 to 0.51.

Keywords: forensic voice comparison, likelihood-ratio, sampling, Bayesian method.

1. INTRODUCTION

Forensic voice comparison (FVC) is a sub-discipline of forensic speech science, which is the application of phonetics, acoustics, signal processing and logic to legal cases [9]. A typical scenario for a FVC case is to compare recordings, one of an unknown offender, and the other of a known suspect typically recorded during the police interview (e.g. in the UK, China) [6] or through wiretaps (e.g. in Germany, China) [12]. The likelihood ratio (LR) framework has been extensively promoted in recent years [9,15,21,23]. The LR approach involves evaluating the similarity of the speech patterns in the disputed and known samples and assessing their typicality against a relevant population [8,10]. The outcome, which can be expressed using a numerical or verbal LR, is a measure of the strength of the evidence under competing propositions of the prosecution and defence (see further [10,14]). Calculating a LR involves two stages: (1) feature-to-score, and (2) score-to-LR. In stage one, measurements or observations from the training and test data are extracted to calculate SS and DS training and test scores. In stage two, training scores are used to generate coefficients [4] that are applied to test scores.

It is important to evaluate system performance (i.e. its ability to separate same and different speaker samples) and this is widely done using the log LR cost function (C_{llr}) [5]. The lower the C_{llr} the more accurate the system is. The term *system* here refers to “a set of procedures and databases that are used to compare two samples, one of known sample and one of disputed sample, and produce a LR” [16]. Evaluating system performance is often carried out by taking a group of speakers (e.g. 60 speakers; [9]) and dividing them equally into training, test and background sets. The system is then trained and tested by using these three sets of speakers. Previous studies have shown that system performance varies when using demographically matched or mismatched speakers for the background population [8]. Different variables also yield different system accuracy [9,10,15,21,23]. However, most of these previous studies have only carried out the experiment once, with one configuration of speakers in each dataset. Relatively little is known about how stable system performance is if different arrangements of speakers are used.

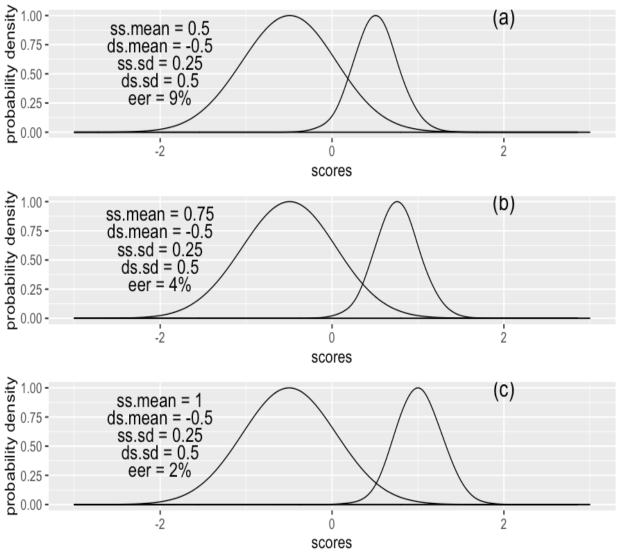
In [22], we conducted a study using spontaneous Cantonese speech from 64 speakers to explore the effect of replicating an experiment multiple times, i.e. by sampling different groups of training, test and background speakers from a relevant population. System performance (C_{llr}s) varied from 0.29 to 1.15 when using different configurations of training, test and background speakers. However, because this study used spontaneous speech, the variability in system stability may have been caused by factors such as number of speakers and tokens used, channel mismatch, and recording qualities of different speakers. Therefore, in order to test purely the effect of the speakers used in each set, it is important to use highly controlled input data, rather than data derived from naturalistic speech. In the present study, same speaker (SS) and different speaker (DS) scores were simulated to address two questions. First, how is system stability is affected by sampling, i.e. does the system have a more stable performance if a different set of training, or test, or training and test speakers are used? Second, do some variables provide more or less stable LR output according to the specific sample of speakers used?

2. METHOD

2.1 Data simulation

To generate controlled data, we simulated comparison scores for the training and test sets. In this way, the study does not focus on the feature-to-score stage and the make-up of the background set, although these are important issues that we explore in [22]. The scores were simulated under an assumption of normality. It is worth noting that scores from real speech data are often not normally distributed. However, normal distributions are used here for the sake of simplicity. Three sets of simulated scores for 1000 speakers were computed using the *rnorm* function in R [1,20], resulting in 1,000 SS and 99,000 DS scores. Panel (a) in Figure 1 shows the distributions of the simulated SS and DS scores, where the mean and standard deviation of SS scores (right) are 0.5 and 0.25, while the mean and standard deviation of DS scores (left) are -0.5 and 0.5. The DS scores have a higher standard deviation because in all FVC with multivariate LRs, the non-target values have a wider spread than the target values. In panels (b) and (c), the standard deviation of SS and DS scores and the mean of DS scores were kept the same, but the mean of SS scores was increased to 0.75 and 1. The three different datasets each had different equal error rates (EER), in order to mimic variables with different speaker-discriminatory power. The data in panel (c) (EER = 2%) has the best speaker-discriminatory power; cf. panel (b) (EER = 4%) and panel (a) (EER = 9%). This allows us to assess the effect of inherent speaker-discriminatory power on system stability. The three sets of scores were used as the pseudo-datasets for LR computation.

Figure 1: Simulated SS (right) and DS (left) scores with different speaker-discriminatory power.



2.2 LR computation and system evaluation

Since the current study uses simulated scores, only the score-to-LR stage of LR computation is assessed here. Previous research shows that stable LR output can be achieved with 20 or more speakers in each of the training and test data [7]. Therefore, 20 training and test speakers were selected randomly from pseudo-datasets (a), (b) and (c) respectively, which led to 20 SS and 380 DS training and test scores. The training scores were used to generate logistic regression calibration coefficients [4] that were then applied to test scores to produce a set of 20 SS and 380 DS calibrated log LRs. The C_{llr} was calculated to capture the system performance. The same procedure was repeated 100 times by using the *LR calculation and testing in FVC* package [13] in R [2,19]. The overall and interquartile range (IQR) of C_{llr}s are used to evaluate system stability.

3. EXPERIMENT

Three experiments were carried out with pre-defined sampling rules. The scores were sampled from pseudo-datasets (a), (b) and (c) in each experiment.

3.1 Expt. 1: Sampling training & test scores.

Different sets of scores were randomly sampled for both training and test data in each replication to explore the effect of score-sampling on system stability, and whether some variables (represented by EER conditions) produce more or less stable systems according to different samples of scores used.

3.2 Expt. 2: Only sampling training scores.

Different sets of scores were randomly sampled for the training speakers while keeping the test scores fixed in each replication. This aims to explore the sensitivity of training data to different speakers with regard to the speaker-discriminatory power of the variable. This allows us to explore whether it matters which training speakers we if the variable has a higher speaker-discriminatory power, i.e. lower EER.

3.3 Expt. 3: Only sampling test scores.

Different sets of scores were randomly sampled for the test data while the training scores were fixed in each replication. This explores the sensitivity of test data to different speakers and the feasibility of using the same LR-based FVC system for multiple cases.

4. RESULTS

4.1. Experiment 1

Figure 2 shows the variation in C_{llr}s by sampling different sets of training and test scores. (a), (b) and (c) indicate the pseudo-dataset that scores were sampled from. Overall C_{llr} ranges from 0.23 to 0.54, 0.10 to 0.42 and 0.03 to 0.51 for sets (a), (b) and (c) respectively. Figure 2 shows firstly that the system stability varies considerably if different sets of SS and DS scores are used in each replication. Furthermore, sets (a), (b) and (c) yielded different system stabilities. Set (c) yielded a lower IQR than sets (a) and (b) (Table 1). However, set (c) also yielded a higher C_{llr} overall range (OR) and more outliers than sets (a) and (b). The results show that score-sampling has a marked effect on system stability regardless of discriminatory power of the feature being used.

Figure 2: Variation of C_{llr}s by sampling training and test scores from pseudo-datasets (a), (b) and (c).

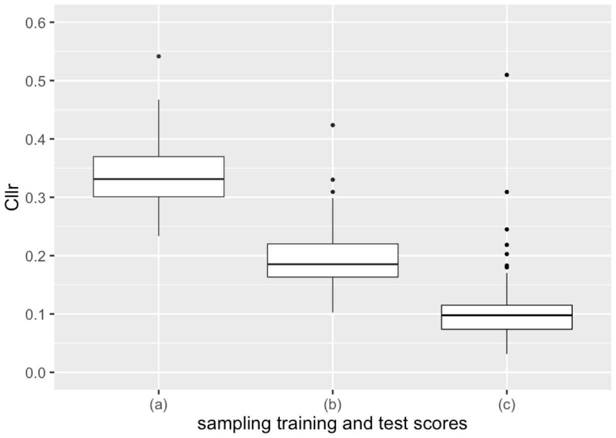


Table 1: C_{llr} minimum, 1st quartile, median, 3rd quartile, maximum, IQR and OR of sets (a), (b) and (c) in experiment 1.

C_{llr}	(a)	(b)	(c)
Min.	0.23	0.10	0.03
1st Qu.	0.30	0.16	0.07
Median	0.33	0.19	0.10
3rd Qu.	0.37	0.22	0.12
Max.	0.54	0.42	0.51
IQR	0.07	0.06	0.05
OR	0.31	0.32	0.48

Experiments 2 and 3 explore the effects found in Experiment 1 in more details, to identify whether the training or test data is more important.

4.2. Experiment 2

Figure 3 shows the variation in C_{llr}s by sampling different sets of training scores in each replication. One predictable pattern emerges, namely that the further apart the distributions of SS and DS scores, the lower the C_{llr} mean and median. All three sets yielded the same C_{llr} IQR (Table 2), which indicates that variables with higher speaker-discriminatory power do not necessarily yield a higher system stability when varying the training speakers. However, more outliers are produced when the distributions of SS and DS scores are further apart from each other, which makes the overall C_{llr} range of set (c) higher than (a) and (b).

Figure 3: Variation of C_{llr}s by sampling training scores from pseudo-datasets (a), (b) and (c).

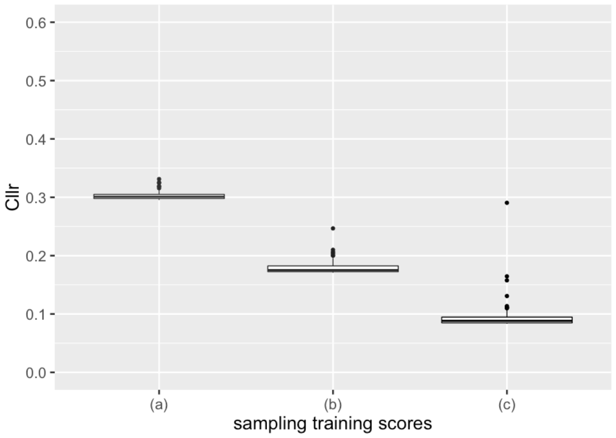


Table 2: C_{llr} minimum, 1st quartile, median, 3rd quartile, maximum, IQR and OR of sets (a), (b) and (c) in experiment 2.

C_{llr}	(a)	(b)	(c)
Min.	0.30	0.17	0.08
1st Qu.	0.30	0.17	0.08
Median	0.3	0.17	0.09
3rd Qu.	0.31	0.18	0.09
Max.	0.33	0.25	0.29
IQR	0.01	0.01	0.01
OR	0.03	0.08	0.21

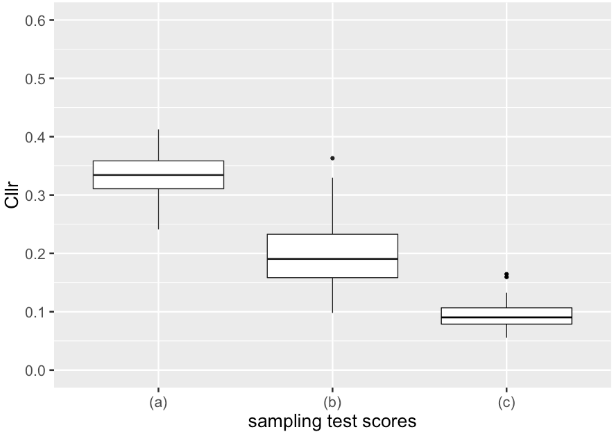
4.3. Experiment 3

Figure 4 shows variation in C_{llr}s when sampling different sets of test scores in each replication. The overall range and IQR of C_{llr}s in Experiment 3 yielded a different pattern from Experiment 2. The C_{llr}s of set (b) range from 0.1 to 0.36 (Table 3) and the IQR is 0.08. These are higher than those of sets (a) and (c). Scores sampled from pseudo-dataset (c) yielded the lowest overall C_{llr} range (0.1) and IQR (0.03), which suggests that it is feasible to use the same LR-based FVC system for multiple FVC cases. However, a comparison between sets (a) and (b) shows a different pattern, suggesting that a variable with a higher speaker-discriminatory power does not always yield higher system stability if different test speakers are used.

Table 3: C_{llr} minimum, 1st quartile, median, 3rd quartile, maximum, IQR and OR of sets (a), (b) and (c) in experiment 3.

C_{llr}	(a)	(b)	(c)
Min.	0.24	0.10	0.06
1st Qu.	0.31	0.15	0.08
Median	0.33	0.19	0.09
3rd Qu.	0.36	0.23	0.11
Max.	0.41	0.36	0.16
IQR	0.05	0.08	0.03
OR	0.17	0.26	0.10

Figure 4: Variation in C_{llr}s by sampling test scores.



5. DISCUSSION

The results from the three experiments showed that score sampling has different effects on system stability.

Experiment 1 shows that system accuracy is not necessarily positively correlated with system stability, because the further away the SS and DS distributions (lower EER) are, the more outliers the system yields.

Experiment 2 shows a similar pattern comparing to Experiment 1. The system stability in Experiment 2 is much higher than that in Experiment 1, and Figure 3 suggests that varying training scores has a limited effect on system stability. However, this may also be due to the fact that when training scores are sampled, the same set of test scores are being used in each replication. Therefore, the effect of sampling variability is overall reduced, which in turn improves the system stability.

In Experiment 3, it was the same calibration coefficients (i.e. same set of training scores) used in each replication. The results suggest that sampling test scores has more effect on the system stability than sampling training scores. Set (c) yielded more outliers than sets (a) and (b), which suggests a low feasibility for using the same LR-based FVC system for multiple real cases even when the variables have a higher speaker-discriminatory power (lower EER). However, set (c) also yielded the lowest OR and IQR, which may suggest that the system starts to yield stable performance when a certain accuracy level is achieved, and a well-defined training data is used [8].

Potential solutions could be developed to deal with the system variability caused by score sampling. First, different calibration methods proposed in [18] might offer a solution to improve system stability. Second, the effect of score sampling on system stability can be explored for different types of distributions by taking skewness and kurtosis from real speech data into consideration [3]. Third, similar to [8], a well-defined training data set may be needed to improve the overall system stability.

6. CONCLUSION

The current study used simulated data to explore the effect of score sampling on system stability. The results reinforce the underlying uncertainty in data-driven FVC studies. The results have a number of implications for both LR-based FVC and phonetic studies in general. Firstly, it is necessary to capture both system accuracy and stability rather than reporting one single LR value in LR-based FVC cases. Secondly, variability in source data causes the system performance to vary to different extents regardless of the speaker-discriminatory power of the variables being used; variables with higher speaker-discriminatory power do not necessarily yield higher system stability. Thirdly, it is essential to replicate experiments multiple times. Otherwise, the results might be misleading if they are used as evidence, with potentially serious consequences for justice.

7. ACKNOWLEDGEMENT

This research was funded in part by the Department of Language and Linguistic Science, University of York. We are also grateful to Justin Lo for the R scripts which facilitated data analysis.

8. REFERENCES

[1] Ahrens, J. H., Dieter, U. (1973). Extensions of Forsythe's method for random sampling from the normal distribution. *Mathematics of Computation,* 27, 927-937.

[2] Aitken, C. 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.

[3] Ali, T., Spreeuwers, L., Veldhuis, R., & Meuwly, D. (2015). Sampling variability in forensic likelihood-ratio computation: A simulation study. *Science & Justice*, 55(6), 499-508.

[4] Brümmer, N. et al. (2007) Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing, 15*, pp. 2072-2084.

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

[6] Home Office. (2003). Criminal Justice Act (Chapter 44). Her Majesty's Stationery Office.

[7] Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15-29.

[8] Hughes, V., Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, *66*, 218-230.

[9] Hughes, V., Foulkes, P., Wood, S. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, 99-132.

[10] Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2(4), 671-711.

[11] 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 & the Law*, *16*(1), 91-111.

[12] Liu, X. M. (2006). 刑事侦查程序理论与改革研究 (Criminal investigation theory and reform). China Legal Publishing House.

[13] Lo, J. (2018). fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison [unpublished R package] version 0.1.0.

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

[15] Morrison, G. S. (2009). 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.

[16] Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*(2), 173-197.

[17] Morrison, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science & Justice*, *56*(5), 371-373.

[18] Morrison, G. S., Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200-218.

[19] Morrison, G.S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. [Software].

[20] R Core Team (2018). R: A language and environment for statistical computing. R Foundatin for statistical Computing, Vienna, Austria. http://www.R-project.org/

[21] Rose, P., Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Odyssey 2016*, 326-333.

[22] Wang, B. X., Hughes, V., Foulkes, P. (2018) A preliminary investigation of the effect of speaker randomisation in likelihood-ratio based forensic voice comparison. *IAFPA 2018*. University of Huddersfield, 125-125.

[23] Zhang, C., Morrison, G. S., Thiruvaran, T. (2011). Forensic voice comparison using Chinese/iau/. *Procceedings of the 17th International Congress of Phonetic Sciences.* Hong Kong, City University of Hong Kong.