A preliminary investigation of speaker randomisation in likelihood-ratio based forensic voice comparison

Bruce Xiao Wang, Vincent Hughes, Paul Foulkes
Department of Language and Linguistic Science, University of York
xw961/vincent.hughes/paul.foulkes@york.ac.uk

Please indicate here whether you prefer:
'work in progress' session poster

Please indicate here whether your paper is eligible for the 'Best Student Paper Award': **Yes**

The likelihood-ratio (LR) framework has been employed in many forensic voice comparison studies to test the speaker-specificity of individual vowels and phonetic sequences (e.g. Morrison, 2009; Zhang, Morrison & Thiruvaran, 2011; Rose & Wang, 2016). Other studies have looked at the effects of sample size on LR outputs (e.g. Hughes, 2014, Hughes & Foulkes 2015). However, few studies have considered the variability in system performance as a function of the make-up (rather than size) of the training, test, and reference data sets used in LR-based testing. Typically, FVC studies select a group of speakers to build the background model and run the experiment. One key question is whether this sample of speakers adequately represents the population, or whether a different selection of people would yield significantly different results. A related question is how the sample is split to create training, testing and background sets in order to assess system reliability. The current study addresses this issue by running multiple replications of the same speaker-discrimination experiment to assess the stability of the log_LR cost (C_{llr}; Brümmer and du Preez, 2006) to randomisation of the speakers in the training, testing and background datasets.

Telephone call recordings of over 70 male Cantonese speakers from IARPA babel Cantonese language pack were used (Andrus et al., 2016). Each recording yields approximately 6-8 minutes' net speech. The Cantonese sentence final particle (SFP) /a/ '啊 ah' and one disyllabic word /hea/ '係啊 yes' were chosen as target variables. Tokens were manually segmented and labelled by using a TextGrid in Praat. F1 and F2 values were extracted automatically at 10% intervals across the full vocalic portion of each token using a script (F3 was not available due to recording quality). The raw data were then further corrected manually and outliers were identified using z-scores. The final dataset contains 67 speakers for /a/ and 63 for /hea/. The average number of tokens per speaker is 11 for /a/ and 9 for /hea/ (min. 4 tokens per speaker). Quadratic and cubic polynomials were fitted to the /a/ and /hea/ formant trajectories respectively and the coefficients used as input to compute multivariate kernel density LR-like scores (using Morrison's 2007 script). The first 20 speakers were then used for training data, the second 20 for testing data, and the rest for reference data for /a/ (27 speakers) and /hea/ (23). Scores for the training data were used to generate logistic regression calibration coefficients (Brümmer et al., 2007) that were then applied to the test scores to produce a set of 20 same-speaker (SS) and 190 different-speaker (DS) calibrated log LRs (each DS pair was only used once). The calibrated C_{llr} was calculated to capture the system performance. The same procedure was then replicated 20 times using randomly assigned groups of speakers. Speakers were randomised by using list randomiser (random.org/lists/).

The left panel in Figure 1 illustrates the C_{llr} values of the 20 replications, while the right panel gives the boxplot. The whiskers in the boxplot represent the most extreme data. The box represents the interquartile range and the outliers are marked by the red crosses. Figure 1 shows that the overall performance is moderate, which is probably due to low speaker numbers, poor recording quality and lack of F3s. /hea/ has a better and more stable performance than /a/ in terms of C_{llr} values. However, speaker randomisation results in highly variable system performance. Cllr results vary from 0.74 to

1.14 for /a/, 0.5 to 0.94 for /hea/. These results suggest it is essential to apply such randomisation in speaker discrimination studies to assess the sensitivity of output to the make-up of the datasets used, and thus to guard against LR results that might be especially strong or weak simply because of the random distribution of speakers in system testing. Ongoing work explores the generalisability of these findings, applying the randomisation method on other data sets including hesitation markers in English (Hughes, Wood and Foulkes, 2016).

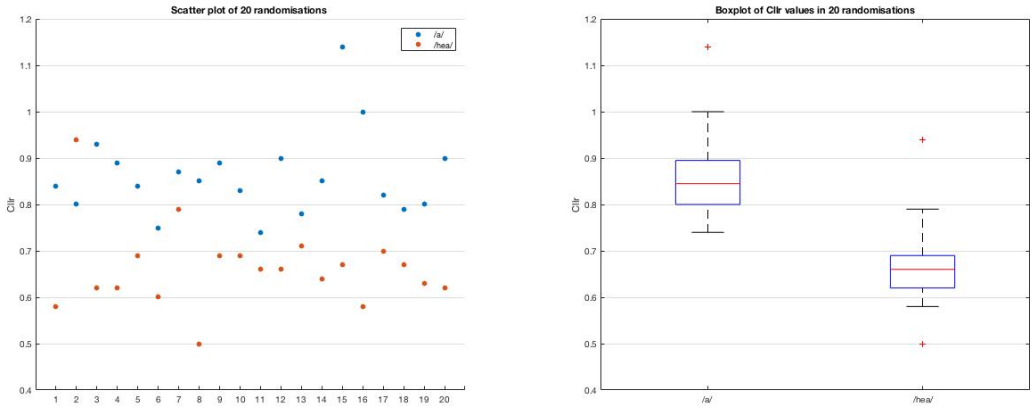

Figure 1. Scatter(left) and boxplot(right) of Cllr values of /a/ and /hea/.

References

Andrus, Tony, et al. (2016) IARPA Babel Cantonese Language Pack IARPA-babel101b-v0.4c LDC2016S02. Web Download. Philadelphia: Linguistic Data Consortium.

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

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, 2072-2084.

Hughes, V. (2014). The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison (Doctoral dissertation), University of York.

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.

Hughes, V., Wood, S. & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, **23(1)**: 99-132.

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.

Morrison, G.S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. http://geoff-morrison.net/#MVKD (accessed 10th Feb 2017).

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.

Zhang, C. Morrison, G. S., & Thiruvaran, T. (2011, August). Forensic voice comparison using Chinese/iau/. In Proceedings of the 17th International Congress of Phonetic Sciences (Vol. 17, p. 21). City University of Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics.