How does the perceptual similarity of the relevant population to a questioned speaker affect the likelihood ratio?

Linda Gerlach[1,2]*, Tom Coy*[2]*, Finnian Kelly*[2]*, Kirsty McDougall*[1]*, and Anil Alexander*[2]

[1]*Theoretical and Applied Linguistics Section, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, Cambridge, UK.*

{lg589|kem37}@cam.ac.uk

[2]*Oxford Wave Research, Oxford, UK.*

{linda|tom.coy|finnian|anil}@oxfordwaveresearch.com

The selection of an appropriate relevant population is a primary concern of experts undertaking forensic speaker recognition (FSR) using a likelihood ratio (LR) framework. In FSR, an assessment of the likelihood that a known (suspect) and a questioned (offender) sample originated from the same or different speakers is formed by considering both the similarity and typicality of the features analysed. While similarity pertains to the features analysed within the given samples, typicality can only be addressed by taking into account the distribution of these features in a bigger population, i.e. the relevant population. Ideally a relevant population is matched to the offender sample in terms of speaker sex, language, accent, channel conditions, recording device, socio-demographic factors and so on (e.g. Hughes 2014, van der Vloed et al. 2020). Research to establish how tightly specified these characteristics need to be in order to form a satisfactory relevant population is ongoing. As an alternative strategy, one would expect that an appropriate relevant population could be determined by using lay-listener judgements of the perceived similarity of voices within a database of speakers (e.g. Morrison et al. 2012). The difficulty in selecting an adequate relevant population based on perceived voiced similarity lies in the fact that many speakers are required to form a relevant population and due to time constraints in forensic casework it is often not logistically possible to collect perceptual similarity judgements from listeners. Further consideration must be given to the effect of choosing very similar speakers to the questioned speaker for the relevant population on the LRs and any bias that may be introduced. Previous research by Gerlach et al. (2020) revealed a significant correlation between listeners' judgements of voice similarity and similarity scores obtained using an automatic speaker recognition (ASR) system that took into account phonetic features linked to perceived voice similarity. This paper explores the role of perceived voice similarity in selecting the relevant population using an ASR system based on auto-phonetic (automatically extracted phonetic) features (Kelly et al. 2019). The adequacy of the selected relevant populations is assessed with regard to the resulting LRs and Cllrs.

This experiment makes use of the GBR-ENG database (2019), a database containing English speech from 600 male and female speakers in 6000 telephone recordings in landline or mobile conditions. Recordings were made across three regions in England (North, Midlands, South), and were each 3-6 minutes in duration. The landline recordings were split into a test set, which was used as mock case files, and a relevant population superset containing recordings from all three regions, from which potential relevant populations were selected.

15 male speakers (5 each from the three regions in England) were chosen randomly from the database. Two landline recordings per speaker were used as mock case files to build 15 same-speaker (SS) comparisons and 210 different-speaker (DS) comparisons. The mock offender samples were compared against the relevant population superset using an auto-

phonetic ASR system in order to choose, for N = 30, 40, …, 100, a) the top N most similar speakers in terms of 'perceived' voice similarity for each ("top"), b) the bottom N most dissimilar speakers ("bottom"), and c) N random speakers ("random"). These selections made up three different relevant populations which were used to evaluate the impact of relevant population speaker similarity on the LRs. The FSR was conducted using an ASR system based on spectral features. LRs were calculated by calibrating the comparison scores within the ASR system (e.g. van der Vloed et al. 2020) using each selected relevant population, and Cllrs were retrieved using Bio-Metrics 1.8 (2019). Preliminary results for an individual mock offender are displayed in Figure 1 and overall show that the absolute value of the average LRs for SS and DS comparisons - and as such the strength of evidence - increases when using a "random" selection of relevant population speakers instead of the "bottom" selection, and increases further when applying the "top" selection of relevant population speakers. The individuals' LRs were compared overall using Cllrs. The lowest (and thus best) Cllrs resulted from using the "top" selection of relevant population speakers and the highest Cllrs from using the "bottom" selection. Overall, the results regarding Cllrs and the LRs of the exemplary mock case indicate that a relevant population containing the perceptually most similar speakers to the questioned speaker could provide better discrimination between SS and DS LRs than a selection of random or least similar speakers.

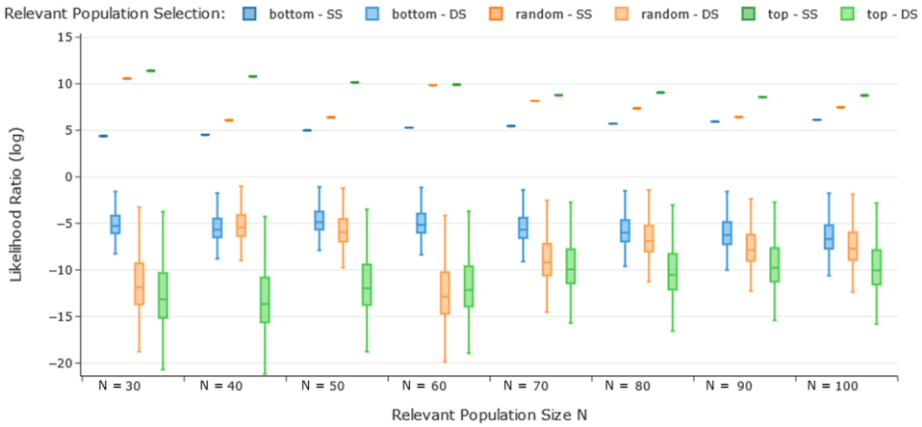

Figure 1 Boxplot of LRs for an individual mock offender's SS and DS comparisons for bottom, random, and top selections of relevant population speakers for different relevant population sizes. Only one LR per SS comparison was available.

References

Bio-Metrics 1.8 performance metrics software (2019), Oxford Wave Research Ltd., https://www.oxfordwaveresearch.com/products/bio-metrics, accessed 31.03.2021.

GBR-ENG database (2019). A telephonic speech database collected for the UK Government for evaluating speech technologies. Further details on application.

Gerlach, L., McDougall, K., Kelly, F., Alexander, A., & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication* 124: 85-95.

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.

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In *Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.*

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. In *Proc. Odyssey Speaker and Language Recognition Workshop.* 62-77.

van der Vloed, D., Kelly, F., & Alexander, A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA: A forensically realistic database. In *Proc. Odyssey Speaker and Language Recognition Workshop*. 402-407.