The effect of background data selection on the strength of evidence

David van der Vloed[1], *Anil Alexander*[2] *and Finnian Kelly*[2]
[1]*Speech and Audio Research, Netherlands Forensic Institute, The Hague, The Netherlands*
d.vandervloed@nfi.nl
[2]*Oxford Wave Research, Oxford, UK*
{anil|finnian}@oxfordwaveresearch.com

In order to estimate likelihood ratios (LRs) or the 'strength of evidence' in forensic speaker recognition cases, the expert must select relevant and representative background data to estimate distributions for same and different speaker comparisons within the conditions of case. These forensic speaker comparisons frequently involve the use of commercially available automatic speaker comparison systems to obtain a case score from the case data (e.g., a suspected speaker's voice recording and a questioned recording), in addition to relevant background data. The use of such background data is only valid if the data selected is adequately representative of the case data. For instance, in a case with a questioned telephone recording and a police interview recording that was recorded on a close microphone, the case score should be considered within the context of same and different speaker distributions obtained using telephone recordings compared to interview recordings. In this study we will use the forensically-relevant FRIDA database (van der Vloed et al., 2018) to create mock cases and to select relevant background data. We will vary the background data and observe any effects on the LRs. The FRIDA database is particularly useful because the recordings in different conditions in this database contain the same speech content, and therefore all differences in LRs will be solely due to the difference in recording conditions. All comparisons will be performed by an x-vector version of VOCALISE (Kelly et al. 2019). The results from this work will help to establish the degree of freedom the practitioner has in choosing representative background data. We will continue this study by involving a test of representativeness proposed in Alexander (2005).

Data and experiment

Six mock cases were selected from FRIDA data, three involving a same speaker comparison and three involving a different speaker comparison. These three cases were chosen from a larger set of comparisons to represent high, medium and low case scores for same and different speaker comparisons. All six mock cases were a comparison between an 'offender' telephone intercept (d5) and a 'suspect' close microphone recording (d2). Background data was selected from 90 speakers from the FRIDA database. The background data representing the offender recording was always d5, and the data representing the suspect recording was varied between d2 (the correct recording device), two other close microphones (d1 and d3), a far microphone (d4) and telephone intercept (d5). d1 and d3 represent a small mismatch with d2; d4 and d5 represent a large mismatch with d2. All recordings used were edited to contain exactly 30 seconds of net

speech, spontaneously spoken in telephone conversations by Dutch speaking males. Case scores were obtained for all six mock cases, and same speaker and different speaker score distributions were produced for each of the five background data sets. All comparisons were performed by VOCALISE without the use of reference normalization or an adaptation cohort. The score distributions were then used to estimate a LR for each case score using BioMetrics 1.6 (2017).

Table 1. LRs using correct (d2) and incorrect suspect background data (d1, d3, d4, d5).

Same source mock cases	***LR d2***	*LR d1*	*LR d3*	*LR d4*	*LR d5*
Low score	**39.4**	34.5	36.0	77.5	7.02
Medium score	**315**	314	327	1625	48.7
High score	**4164**	5894	8560	99030	368
Different source mock cases					
Low score	**$^{1/}$26051**	$^{1/}$28318	$^{1/}$32876	$^{1/}$51272	$< 10^{-10}$
Medium score	**$^{1/}$201**	$^{1/}$168	$^{1/}$188	$^{1/}$121	$< 10^{-5}$
High score	**$^{1/}$9.68**	$^{1/}$6.91	$^{1/}$7.82	$^{1/}$3.61	$^{1/}$455

As can be seen in Table 1, the LRs in columns d1, d2 and d3 do not vary much. This indicates that when selecting background data, the selection of the exact type of close microphone may not be too critical. However, choosing a very different recording device to represent a close microphone (columns d4 and d5) often leads to LRs of different orders of magnitude than the LRs produced with the correct background data. This indicates that although LR estimation is robust to selection of approximately matched background data, using a mismatched background database can significantly impact the strength of the evidence.

References

Alexander, A. (2005). *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*, Swiss Federal Institute of Technology, Lausanne

Bio-Metrics 1.6 performance metrics software (2017), Oxford Wave Research Ltd., http://www.oxfordwaveresearch.com/products/bio-metrics

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

van der Vloed, D., Bouten, J., Kelly, F., and Alexander A. (2018). *NFI-FRIDA – Forensically Realistic Inter-Device Audio*, IAFPA 2018