A WYRED connection: x-vectors and forensic speech data

Anil Alexander[1], Finnian Kelly[1], and Erica Gold[2]
[1]Oxford Wave Research, Oxford, UK, [2]University of Huddersfield, UK
{anil|finnian}@oxfordwaveresearch.com, e.gold@hud.ac.uk

Motivation

Standard Southern British English has been studied extensively both in phonetic and automatic speaker recognition using the well-established DyVIS corpus (Nolan et al. 2009). The West Yorkshire Regional English Database (WYRED; Gold et al. 2018) addressed the lacuna in a representative database of West Yorkshire English, and incorporated and further extended the DYVIS corpus collection protocol by collecting non-contemporaneous data, two additional spontaneous speaking tasks, and recruited a larger pool of speakers. An important addition to WYRED is Task 4, which involves a short, two-minute voicemail message; this type of recording is representative of real case-like material, and yet has not previously been tested for robustness and comparability in forensic speaker comparisons.

WYRED also includes additional speaker demographic information, providing a resource for evaluating new methods and algorithms in phonetic and automatic speaker analysis. In this study, we evaluate the performance of the latest generation of speaker recognition algorithms, namely DNN-based 'x-vectors', in comparison to the previous i-vector approach (Ross et al. 2019), particularly in cross-channel forensically-relevant comparisons, where x-vectors have been shown to provide better discrimination (Kelly et al. 2019). We additionally explore how x-vectors may be used for estimating speaker profile information, including biological gender, age ranges, weight ranges, and spoken language. This exploration of speaker profiling is enabled by the rich meta-data, which is unique to the WYRED collection.

Database

WYRED contains 180 male speakers aged 18-30. The data are evenly divided between Bradford, Kirklees, and Wakefield speakers (60 speakers each). All speakers are from the local areas, were native English speakers, from English-speaking households, and did not have any reported speech or hearing impairments. Ethnicity data was not collected, however, the majority, but not all, speakers were White British.

Speakers participated in a set of 4 tasks that were collected over two recording sessions, which were made at least 6 days apart. Each speaker completed the first two tasks in their first recording session, and the second two tasks in the second recording session. Task 1 is a mock police interview with a research assistant in which participants were told that they were involved in the trafficking of heroin, and to avoid telling the truth about their whereabouts on the day of the crime. Task 2 is a phone call to the participants'

accomplice (a different research assistant) where they were asked to confirm their story and what happened in the police interview. In the second session, Task 3 is a paired conversation where participants spoke with an acquaintance they already knew from the area or they were paired with another speaker from their region. Participants were allowed to discuss whatever they wanted, but topic cards were available if required. Finally, Task 4 is a short voicemail message in which each participant asks their (fictional) brother to get rid of any incriminating evidence. All 4 tasks were studio quality recordings, while Task 2 and 4 were also recorded over a landline telephone.

Speaker Recognition Experiment

The WYRED recordings were first pre-processed by using their accompanying Praat TextGrid files to remove all audio content not pertaining to the speaker of interest. All cross-task recording pairings were compared using VOCALISE (Kelly et al., 2019), a forensic speaker recognition system, in x-vector PLDA mode. The resulting EERs ranged from 0.01% (Task 1 studio vs Task 2 studio) to 2.82% (Task 2 tel vs Task 3 studio). A breakdown of EERs for Task 4 is shown in Table 1, demonstrating the effectiveness of x-vectors for speaker recognition on forensically-relevant recordings, within and cross-condition.

	Task 1 studio	**Task 2 studio**	**Task 3 studio**	**Task 2 tel**
Task 4 studio	0.50%	1.06%	0.72%	2.50%
Task 4 tel	1.07%	1.72%	1.47%	2.40%

Table 1: EERs for all cross-task comparisons in WYRED

The type of forensically-relevant recordings contained in Task 4 are not traditionally seen in databases collected for research. However, the results demonstrate that they are well-suited for automatic analysis. Furthermore, the controlled variability in WYRED enables structured speaker recognition evaluations of this kind, and the meta-data further supports the exploration of speaker profiling.

References

Gold, E., Ross S., and Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proceedings of Interspeech.* Hyderabad, India, 2748– 2752.

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.

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1): 31- 57.

Ross, S., Earnshaw, K. & Gold, E., (2019). A cautionary tale for phonetic analysis: the variability of speech between and within recording sessions, *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, p. 3090-3094.