One out of many: A sliding window approach to automatic speaker recognition with multi-speaker files

Linda Gerlach[1], Finnian Kelly[2], and Anil Alexander[2]

[1]*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*
Gerlach8@students.uni-marburg.de

[2]*Oxford Wave Research Ltd., Oxford, United Kingdom*
{finnian|anil}@oxfordwaveresearch.com

Multi-speaker recordings are frequently encountered in forensic speaker recognition cases. A typical first step in processing such recordings is speaker diarisation, which involves the labelling of every speech event in a recording according to the identity of the speaker. Once diarised, speech from a speaker of interest can easily be extracted for subsequent analysis. Diarisation can be a time consuming process often requiring significant manual effort on the part of the practitioner. Manual or semi-automatic diarisation is not feasible when dealing with large numbers of multi-speaker recordings. In this study, we evaluate the performance of state-of-the-art i-vector and x-vector speaker recognition systems when tested with two-speaker recordings, and assess whether it is possible to bypass speaker diarisation by adopting a simple sliding window approach to speaker recognition. Such an approach would allow for efficient speaker-spotting across large multi-speaker speech databases, which could be very useful for law-enforcement applications.

The DyViS (Nolan, 2011) corpus was used as a source of data for our experiments. DyViS Task 1 (interview) recordings were used as a source of two-speaker data. These recordings consist of a conversation between a participant (our speaker of interest) and an interviewer. Each recording channel contains predominantly the speech from either the participant or the interviewer, although there is some bleeding between the channels. We therefore considered three possible types of two-speaker recordings: channel 1 only (containing predominantly speaker of interest, with some bleeding containing the interviewer), channel 2 only (containing predominantly the interviewer, with some bleeding containing the speaker of interest), and both channels merged (containing speech from both speakers). A set of 100 files for each of these scenarios was compared against 100 Task 3 (report) recordings (all of 120s duration) using VOCALISE (Alexander et al., 2016, Kelly et al., 2019) i-vector and x-vector systems. The resulting EERs are shown in Table 1.

Table 1. EERs for Task 1 (channel 1, channel 2, and merged channels) vs Task 3: 100 speakers

Session	*Channel 1*	*Channel 2*	*Merged*
2018A-Adaptable-15F (**i-vector**)	4.85 %	23.25 %	10.32 %
2019A-Beta-RC1 (**x-vector**)	2.78 %	25.54 %	11.45 %

It is clear that where there is a significant amount of speech from the interviewer (channel 2 and merged), the performance drastically decreases relative to the baseline EER with recordings containing predominantly the speaker of interest. We therefore propose a simple but effective approach to dealing with multi-speaker recordings without preprocessing using manual or automatic diarisation. Our approach requires multiple comparisons per-file, and since we wished to evaluate several permutations of window sizes for both i-vector and x-vector systems, we restricted this experiment to a subset of 10 files from DyViS Task 1 and Task 3. The subset of 10 merged-channel files from Task 1 was selected and split into chunks with varying window lengths of 70, 60, 50, 45, 40, 30, 20, 15, 10, and 5 seconds, using a fixed slide of 5 seconds. For each window length, all of the chunks for each merged Task 1 file were compared against the 10 Task 3 files from the same set

of speakers. One comparison score was obtained for each Task 1 vs Task 3 comparison by taking the maximum score across all of the chunks in the file. The resulting EERs are displayed in Figure 1.

Table 2. EERs for Task 1 (channel 1, channel 2, merged channels, and a 5 second sliding window approach) vs Task 3: 10 speakers

Session	*Channel 1*	*Channel 2*	*Merged*	*Merged Sliding (5 s)*
2018A-Adaptable-15F (**i-vector**)	2.50 %	18.16 %	12.50 %	4.17 %
2019A-Beta-RC1 (**x-vector**)	1.82 %	21.58%	4.00 %	0.00 %

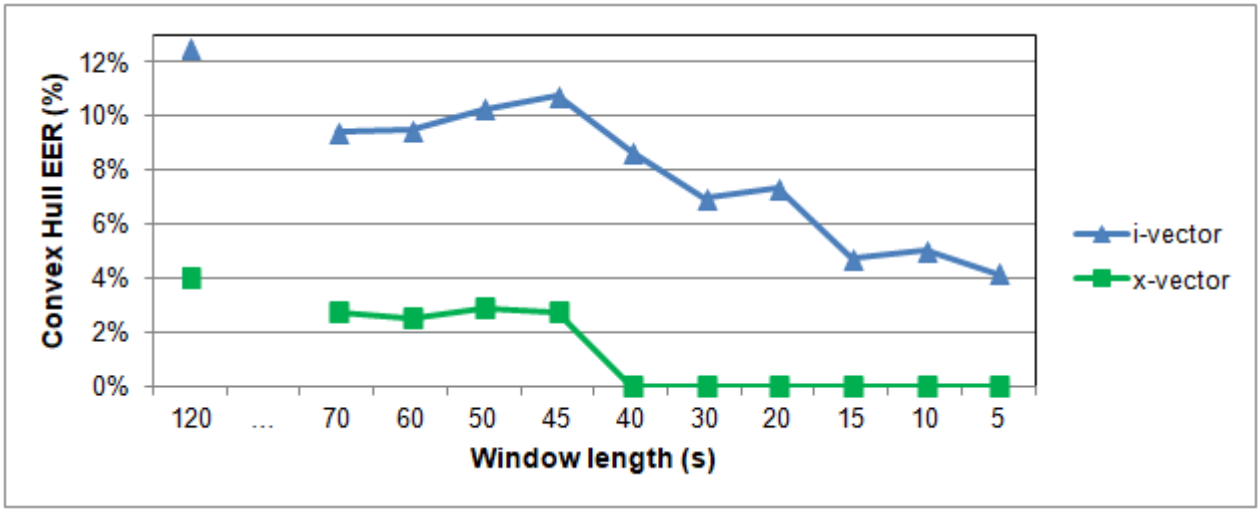

Figure 1: EERs for two-speaker (merged DyViS Task 1 files) comparisons with i-vectors and x-vectors at varying window lengths, with a fixed slide of 5 seconds.

The EERs for both systems generally follow a decreasing trend as the duration of the sliding window decreases; at a window length of 5 seconds, the EER for i-vector and x-vector systems is 4.17% and 0% respectively, a reduction from 12.5% and 4% at a full duration of 120 seconds. The x-vector system clearly outperforms the i-vector system, dropping to an EER of 0% by 40 seconds. These results support the use of a sliding window approach to speaker recognition of two-speaker files when diarisation is not feasible. We will present further results to demonstrate the application of this approach to two-speaker files from the forensically-relevant FRIDA (van der Vloed et al., 2018) database.

References

Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. (2016). *VOCALISE: A Forensic Automatic Speaker Recognition System supporting Spectral, Phonetic, and User-Provided Features.* Odyssey 2016.

Nolan, F. (2011). *Dynamic Variability in Speech: a Forensic Phonetic Study of British English, 2006-2007*. [data collection]. UK Data Service. SN: 6790, http://doi.org/10.5255/UKDA-SN-6790-1

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.*