

Discovery and retrieval of speakers from large unlabelled datasets using scalable clustering

Thomas Coy, Finnian Kelly and Anil Alexander
Oxford Wave Research Ltd, Oxford, United Kingdom
{finnian|tom.coy|anil}@oxfordwaveresearch.com

In digital forensic investigations containing large quantities of multimedia files it may be necessary to rapidly triage hundreds or thousands of unlabelled files, to either find previously unknown speakers of interest or to search for specific speakers. Without any indication as to how many speakers are in each file, who those speakers are, or whether speakers reappear elsewhere in the unlabelled data, this could become exceedingly challenging for both manual or computational methods. We propose a two-tier clustering-based approach to address these issues. The first stage is to run within-file clustering to create a cluster database. This then enables:

1. Retrieval - compare a single-speaker reference sample against the database
2. Discovery - second-level clustering to group within-file clusters that contain the same speaker

Naive clustering involves comparing every identified cluster in each file against all clusters in the database. This has $O(n^2)$ complexity that scales poorly as the number of files increases, making clustering with thousands of files computationally prohibitive. Here we propose a scalable clustering that sidesteps this issue.

Pre-processing includes the removal of audio of inadequate quality and regions where speakers are overlapping. The remaining audio is then segmented and x-vectors extracted, before running agglomerative hierarchical clustering to estimate how many speakers are in a file and where they are speaking, creating a within-file cluster database.

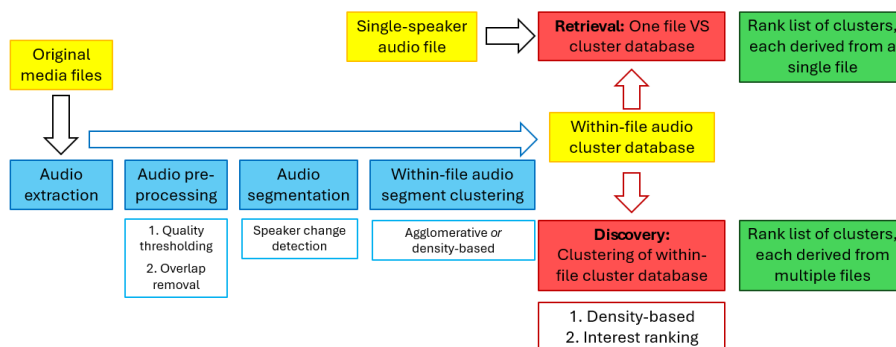


Figure 1. A flowchart showing the sequence of clustering processes.

For retrieval, VOCALISE (Kelly et al, 2019) is used to compare a reference sample against the database. For discovery, a density-based clustering algorithm is used to find groups of within-file clusters belonging to the same speaker. Each discovery group undergoes an all versus all

VOCALISE comparison to generate comparison metrics to rank groups by match strength (i.e. a high mean score suggests a more coherent group).

To evaluate performance, a curated subset of 386 files from VoxCeleb1 (Nagrani et al, 2017) was used. Files contain speech from at least two speakers, only one of which is labelled, and speakers appear in more than one video. There is diversity in recording conditions and speaker population.

Discovery elicited a total of 407 groups containing clusters from 2 or more different files. Of these, 26 groups were determined to be pure (i.e. containing a single, labelled speaker). Pure groups covered 22 of the 129 unique labelled speakers in the curated dataset. Thus, there were several labelled speakers who appeared in more than 2 pure groups. Of the remaining ‘impure’ groups, some were manually inspected and discovered to contain the same unlabelled speaker across files. For example, the fourth, blue-highlighted group in Figure 2 appears to contain results from Speakers 68 and 79. This group actually consists of speech from a mutual interviewer who appears in both files. As such the algorithm found links that were unknown to us.

By deploying clustering algorithms, rather than non-scalable, traditional NvN comparisons, thousands of completely unlabelled multimedia files can be rapidly analysed.

File Name	Audio Quality	Group Mean	No. of Original Files	Ind. Mean
speaker029_F_USA_sample001_cluster0002.wav	4	61.61	2	61.61
speaker029_F_USA_sample003_cluster0003.wav	4	61.61	2	61.61
speaker057_F_USA_sample003_cluster0005.wav	3	52.68	2	52.68
speaker068_F_USA_sample004_cluster0002.wav	3	52.68	2	52.68
speaker010_M_USA_sample002_cluster0001.wav	5	41.92	2	41.92
speaker010_M_USA_sample004_cluster0001.wav	5	41.92	2	41.92
speaker068_F_USA_sample002_cluster0002.wav	4	36.64	2	36.64
speaker079_M_USA_sample001_cluster0005.wav	5	36.64	2	36.64
speaker057_F_USA_sample001_cluster0004.wav	4	31.49	4	21.72
speaker057_F_USA_sample002_cluster0004.wav	4	31.49	4	30.08
speaker057_F_USA_sample004_cluster0004.wav	3	31.49	4	36.64
speaker057_F_USA_sample006_cluster0004.wav	3	31.49	4	37.53
speaker002_M_IND_sample001cluster0004.wav	3	28.27	2	28.27
speaker002_M_IND_sample002cluster0003.wav	5	28.27	2	28.27
speaker056_M_USA_sample001_cluster0002.wav	2	27.87	4	38.34
speaker056_M_USA_sample003_cluster0004.wav	2	27.87	4	24.26
speaker056_M_USA_sample004_cluster0004.wav	2	27.87	4	22.83
speaker056_M_USA_sample005_cluster0003.wav	4	27.87	4	26.05
speaker083_F_USA_sample001_cluster0002.wav	5	25.29	3	39.26
speaker083_F_USA_sample002_cluster0004.wav	4	25.29	3	30.81
speaker083_F_USA_sample003_cluster0002.wav	3	25.29	3	5.79
speaker034_F_MEX_sample002_cluster0001.wav	3	24.47	4	-7.90
speaker034_F_MEX_sample003_cluster0004.wav	4	24.47	4	36.86
speaker034_F_MEX_sample005_cluster0002.wav	4	24.47	4	33.79
speaker034_F_MEX_sample006_cluster0003.wav	4	24.47	4	35.15

Figure 2. Discovery results showing groups of within-file clusters that contain the same speaker.

References

- Kelly, F., Forth, O., Kent, S., Gerlach, L. & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Proc. AES International Conference 2019*, Paper 27.
- Nagrani, A., Chung, J. S. & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. *INTERSPEECH*, (2017).
- Ruch, H., Fröhlich, A. & Lim, S. (2023). Clustering a large number of unknown voices. *Proc. International Association for Forensic Phonetics and Acoustics (IAFPA) Conference, Zurich, Switzerland*, 23-24.