Seeking voice twins – an exploration of VoxCeleb using automatic speaker recognition and two clustering methods

Linda Gerlach[1,2]*, Kirsty McDougall*[1]*, Finnian Kelly*[2]*, 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|finnian|anil}@oxfordwaveresearch.com`

Speaker similarity is a highly relevant concept in forensic phonetics, be it for constructing a voice parade fair to all involved parties or to assess the theoretical impact of relevant populations similar to a suspect speaker in forensic speaker recognition. Taking similarity to the extreme, the question arises whether it is possible to find voice twins, i.e. speech recordings originating from different, unrelated speakers that sound extremely similar to one another. Applications of voice twins may be found in earwitness assessment tasks (see Schäfer & Foulkes, 2022) or in medical voice banking when the available audio material of a voice-impaired person is insufficient for personalising a speech-generating device (see e.g. Yamagishi et al., 2012).

Previous studies have indicated that automatically obtained similarity scores based on perceptually relevant acoustic features (i.e. LTF1 to LTF4) are able to approximate, to a certain extent, ratings of perceived voice similarity as judged by listeners (Gerlach et al., 2020, 2021). Using automatically obtained similarity scores, recent research has tried to further concentrate a selection of speakers into more similar subgroups using agglomerative hierarchical clustering (AHC), gaining some general insights regarding clustering of speaker sex and the potential to find very similar sounding speakers in clusters where AHC branches merge early on (Gerlach et al., 2022). However, AHC has the disadvantage of forcing items into clusters and hence may form them even when speakers do not sound particularly similar. Additionally, the study relied on a small selection of 180 speakers with one recording each, thus, increasing the number of speakers as well as the diversity of recordings may improve the chances of discovering voice twins.

The aim of the present study is to expand on Gerlach et al. (2022) using a subset of good quality recordings (n=831, thereof 348 female; 30s minimum net speech, 24dB SNR, 0% clipping) of VoxCeleb (Nagrani et al., 2020), a large, diverse speaker database encompassing multiple recordings per speaker. Two clustering approaches for detecting voice twins will be explored: the previously used AHC, as well as the clustering method DBSCAN (density-based spatial clustering of applications with noise), which allows for items to not belong to a cluster ("noise"). Similarity ratings between the recordings will be obtained using VOCALISE automatic speaker recognition software relying on x-vectors and automatically-extracted phonetic features (LTF1 to LTF4). It is hypothesised that dense clusters containing more than one speaker and multiple files per speaker are possible voice twin candidates. An initial auditory and acoustic assessment of potential voice twins will be conducted, and challenges pertaining to voice similarity assessment and constructing a listener experiment will be discussed.

References

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2021, August). How do automatic speaker recognition systems 'perceive' voice similarity ? Further exploration of the relationship between human and machine voice similarity ratings. *Proceedings of the Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA).*

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2022, April). Selecting similar-sounding speakers for forensic phonetic applications: an exploration of cluster analysis using automatic speaker recognition. *Presented at British Association of Academic Phoneticians (BAAP) Colloquium*.

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. https://doi.org/10.1016/j.specom.2020.08.003

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, *60*, 101027. https://doi.org/10.1016/j.csl.2019.101027

Schäfer, S., & Foulkes, P. (2022, April). Assessing the Individual Voice Recognition Skills of Earwitnesses. *Presented at British Association of Academic Phoneticians (BAAP) Colloquium*.

Yamagishi, J., Veaux, C., King, S., & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, *33*(1), 1–5. https://doi.org/10.1250/ast.33.1