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

Linda Gerlach[1,2] (lg589@cam.ac.uk), Kirsty McDougall[1], Finnian Kelly[2], and Anil Alexander[2]
[1]University of Cambridge, Cambridge, UK
[2]Oxford Wave Research Ltd., Oxford, U.K.

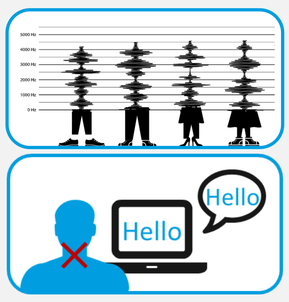
1. Motivation

What are voice twins?

- Located at the extreme end of perceived voice similarity, voice twins may be different, unrelated speakers that sound extremely similar to one another.
- Acoustic features contributing to speaker similarity include F0, LTF1 to LTF4 [1].

Why should we care?

- Voice twins could help in gaining further insights into human perception of speaker individuality.



Research questions:

1. Is it possible to find different, unrelated speakers that sound extremely similar to one another, i.e. voice twins, automatically?
2. Can human listeners successfully differentiate between speech from "voice twins", from the same speakers, and from random speaker pairs?
3. How confident are listeners in their judgements?

2. Proposed methodology

Finding speakers that sound extremely similar:

1. Select perceptually relevant features
2. Model features automatically
3. Calculate similarity estimates for speaker pairs
4. Use these similarity estimates to find subgroups of similar-sounding speakers
5. Explore clustering techniques to find these subgroups of potentially indistinguishable speakers
6. Assess how difficult it is to differentiate between speakers in the subgroups identified using human listeners

3. Speaker databases

- Subsets from three speaker databases: small and controlled to large and diverse
- Selection of good quality recordings:
  - Net speech > 20 s
  - SNR > 18 dB
  - Clipping < 50 %
- Selection of two random files per speaker to model intra- and interspeaker variability

WYRED [2]:
Homogeneous database
180 male speakers of similar age
English language, similar accents
Task 2 (accomplice telephone call)
Task 4 (answerphone message)
Studio quality recordings

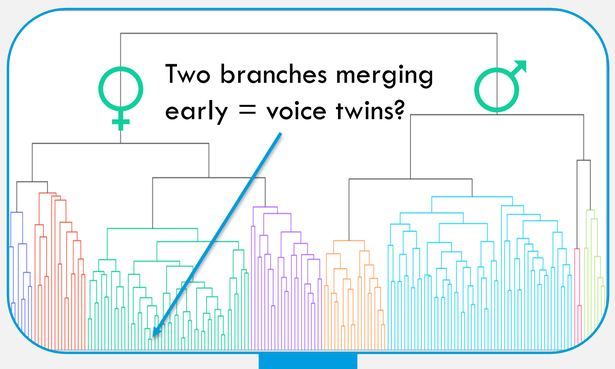
GBR-ENG [3]:
390 speakers (202 female)
English language, various accents
Landline recordings made in GB

VoxCeleb [4]:
Highly diverse database
906 speakers (metadata not available)
Mostly English language, various accents
Interview recordings from YouTube

4. Experiment

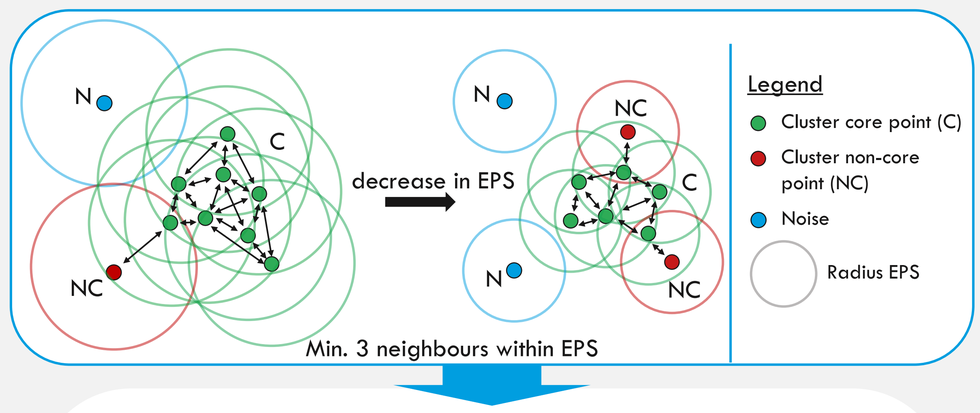
Automatic speaker recognition

- Extraction of perceptually relevant phonetic features LTF1 to LTF4 using x-vectors [5]
- Creation of voice models and calculation of similarity scores using VOCALISE automatic speaker recognition software [6]



Agglomerative hierarchical clustering (AHC) groups recordings hierarchically according to pairwise distances (similarity) [7].

- Advantage:
  - All recordings will form part of a cluster
- Disadvantages:
  - Clusters are determined visually in dendrogram
  - Maximum number of clusters needed as input
- Pilot experiment using a GBRENG subset [8]:
  - 180 unique speaker recordings (90 female)
  - Clusters according to speaker sex
  - Difficult to assess cluster similarity further



DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups recordings within dense areas as defined by a minimum number of neighbouring points within the radius EPS [9].

- Advantages:
  - Recordings that are not close to others are penalised (= noise)
  - Varying the distance measure EPS may result in voice pairs of different degrees of similarity
- Disadvantage:
  - Not all recordings may form part of a cluster
- In this experiment:
  - Male and female speakers processed together
  - Voice twin candidates = clusters containing both recordings of two different speakers

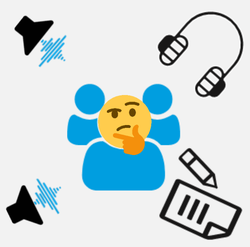
5. Current results

Number of voice twin candidates per database for different distance values (EPS):

EPS	WYRED	GBRENG	VoxCeleb
0.2	1	3 ♀, 1 ♂	1 ♀, 0 ♂
0.25	4	4 ♀, 6 ♂	1 ♀, 1 ♂
0.3	4	2 ♀, 8 ♂	0 ♀, 2 ♂

6. Next: Listener experiment

- Three blocks (one per dataset)
- Same/different judgements, 1-5 confidence scale
- Sample duration: 3-5s
- Comparisons to include:
  - Voice twins
  - Same-speaker pairs
  - Random different-speaker pairs
- Two files per speaker, three samples per recording
- Listeners will judge subsets of the total number of comparisons



Ask to listen to some samples!

7. Hypotheses assessed

1. The automatic method produces voice twin pairs that can be used in listener experiments.
2. Listeners will perform worse when discriminating between speech from voice twins than from random speakers (higher number of false acceptances).
3. Listeners' confidence in their judgements will vary according to the type of comparison in line with [10].

References

[1] McDougall, K. (2021). Ear-Catching versus Eye-Catching? Some Developments and Current Challenges in Earwitness Identification Evidence. Proceedings of XVII AISV (Associazione Italiana Scienze Della Voce) Conference: 'Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications', [Preprint].
[2] Gold, E., Ross, S., & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Paper 0065, 2748-2752.
[3] GBR-ENG database. (2019). *A telephonic speech database collected for the UK Government for evaluating speech technologies.*
[4] Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, *60*, 101027.
[5] 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).
[6] 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.
[7] San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., & Kavanagh, C. (2018). Cluster analysis of voice quality ratings: Identifying groups of perceptually similar speakers. In: Proceedings of the Conference on Phonetics & Phonology in German-speaking countries (P&P 13). pp.173-176.
[8] 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.
[9] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Transactions on Database Systems, 42(3).
[10] Afshan, A., Kreiman, J., & Alwan, A. (2022). Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech. The Journal of the Acoustical Society of America, 151(2), 1393–1403.