Can voice similarity be assessed using an automatic speaker recognition system?

UNIVERSITY OF
CAMBRIDGE

Linda Gerlach^{1,2}, Kirsty McDougall¹, Finnian Kelly², and Anil Alexander²
¹University of Cambridge, Cambridge, UK, ²Oxford Wave Research Ltd., Oxford, U.K.

1. Introduction

- Scenario:** A crime has been committed but the witness has only heard the perpetrator and a recording is not available. A suspect has been determined and a **voice parade** is required.

- A number of foil voices must be selected and a fair comparison for the suspect must be ensured by assessing **perceived voice similarity**.
- Current procedures are time-consuming, involving substantial manual effort including two pre-parade experiments relying on lay listeners to evaluate fairness [1].
- An initial study considering an **automatic speaker recognition** (ASR) system to assess voice similarity of a group of SSBE speakers from a forensically relevant database showed promising results, encouraging further research [2].

2. Objectives

The present study aims to

- Test whether the statistically significant, positive correlation between automatically obtained voice similarity estimates and ratings by lay listeners found in [2] holds for other groups of speakers from the same accent background.
- Test whether the correlation found in [2] also holds for different-accented speakers of English.

3. Speaker databases & groups

	DyViS [3]	YorViS [4]	WYRED [5]
Accent	SSBE	York English	West Yorkshire Englishes (Bradford, Kirklees, Wakefield)
Total No. of speakers	100	20	180 (60 per location)
Gender	Male	Male	Male
Age	18-25	18-25	18-30
Speaker groups for experiment	3 x 15 speakers (D1, D2, D3)	1 x 15 speakers (Y)	1 x 15 speakers Bradford (W1) 1 x 15 speakers Wakefield (W2)
Speaking tasks	1 & 2 (spontaneous speech)	1 & 2 (spontaneous speech)	1 & 2 (spontaneous speech)
Recording quality	Studio	Studio	Studio

4. Methods

Listener experiment:

- Listener ratings were collected as part of the VoiceSim [6] and IVIP [7] projects and made available for this study.
- Paired speaker comparisons using short stimuli (~3 s) were collected for each speaker group (120 comparisons, incl. same-speaker; two samples per speaker).
- Voice similarity ratings from 120 listeners (age 18-40, English L1, no hearing impairments, similar number of males and females; 20 listeners per speaker group) were collected.
- Similarity was judged on a scale from 1 – very similar to 9 – very different.

Automatic experiment:

- The forensic automatic speaker recognition software VOCALISE [8] was applied to automatically obtain voice similarity estimates for the speakers within each speaker group:
 - Two samples per speaker (~4 min each).
 - Speaker modelling based on automatically extracted phonetic features that are considered relevant for perceived voice similarity (incl. F1-F4 and F0).
 - Application of an i-vector framework [9], where an i-vector represents a speaker in a compressed, fixed number of dimensions [8].
 - Probabilistic Linear Discriminant Analysis (PLDA) to obtain comparison scores (here used as voice similarity estimates) for each speaker pair within a speaker group.
- Normalisation of comparison scores regarding their numerical range using Bio-Metrics [10] to ensure comparability across speaker groups.
- Averaging of comparison scores of the same speaker pairs (spk1_smpl1 vs spk2_smpl2 and spk1_smpl2 vs spk2_smpl1) to obtain one score per speaker pair.

Heat map of calibrated, averaged VOCALISE comparison scores for DyViS D1, including same-speaker comparisons on diagonal. Gradient from red (low score) to green (high score).

Evaluation and analysis:

- Calculation of 5% trimmed mean of the listener ratings for each comparison to account for variable use of the rating scale.
- Correlation analysis to explore the degree and direction of the relationship between listener ratings of voice similarity and automatically retrieved scores for the different speaker groups.
- Analysis for different-speaker comparisons only, using
 - Spearman rank correlation (2-tailed)
 - Pearson's correlation (2-tailed) to further assess the linearity of the relationship

5. Results

Scatter plot showing listener ratings and automatically obtained scores for speaker groups DyViS D1, D2, and D3.

Speaker group	Spearman's rho (2-tailed)	Pearson's r (r ²)
DyViS D1	0.484 (p < .001)	0.503 (0.253)
DyViS D2	0.519 (p < .001)	0.502 (0.252)
DyViS D3	0.357 (p < .001)	0.366 (0.134)
WYRED W1	0.380 (p < .001)	0.398 (0.158)
WYRED W2	0.551 (p < .001)	0.631 (0.398)
YorViS Y	0.463 (p < .001)	0.704 (0.500)

Spearman and Pearson correlations of listener ratings and automatically obtained voice similarity estimates for each speaker group. Independent of the accent of a speaker group, the correlations indicate a **statistically significant, positive relationship** between listener ratings and automatic estimates. The linearity of the relationship varies, showing no clear trend with respect to different speaker accents.

6. Conclusion

- The results from this study corroborate previous findings indicating that there is a notable correlation between voice similarity assessments by listeners and by an ASR system which takes into account perceptually relevant phonetic features independent of the accent within the speech samples.
- Future work will consider
 - Whether this relationship can also be found when using the more common feature extraction approach relying on spectral features to model the speakers, and
 - Whether applying the latest ASR approach using x-vectors based on Deep Neural Networks to model a speaker and subsequently assessing voice similarity yields similar correlations, and
 - Aim to establish score thresholds to enable the fair selection of sufficiently similar voices for, e.g., voice parades.

References

- [1] de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice Lineups: A Practical Guide. 18th International Congress of Phonetic Sciences, edited by The Scottish Consortium for ICPhS 2015, Paper number 0598, 1-5.
- [2] 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.
- [3] Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic research. *International Journal of Speech, Language, and the Law*, 16(1), 31-57.
- [4] McDougall, K., Duckworth, M., & Hudson, T. (2015). Individual and group variation in disfluency features: A cross-accent investigation. 18th International Congress of Phonetic Sciences, edited by The Scottish Consortium for ICPhS 2015, Paper number 0306, 1-5.
- [5] 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. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, September 2018, Paper number 0065, 2748-2752.
- [6] <https://www.phonetics.mml.com.ac.uk/research/voicesim/>, last accessed 10.06.2021.
- [7] <https://www.phonetics.mml.com.ac.uk/ivip/overview/>, last accessed 10.06.2021.
- [8] Alexander, A., Forth, O., Atreya, A. A., & Kelly, F. (2016). VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features. *Odyssey 2016 Show & Tell*, Paper number 88, 1-4.
- [9] Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J., & Åkesson, J. (2016). Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, September 2016, Paper number 2018, 1567-1568.
- [10] Bio-Metrics 1.8 performance metrics software, Oxford Wave Research Ltd., <https://www.oxfordwaveresearch.com/products/bio-metrics/>, last accessed 10.06.2021.