How do automatic speaker recognition systems 'perceive' voice similarity? Further exploration of the relationship between human and machine voice similarity ratings

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

Perceived voice similarity plays a crucial role in the construction of voice parades, where foil voices must be selected in a way that is fair to witness and suspect. Current procedures are limited in that they involve extensive manual effort to collect and process voice similarity judgements of naïve listeners (e.g. McDougall 2013). Automating steps in the selection of similar voices could speed up the construction of voice parades, reducing the cost and effort required and potentially resulting in more parades being conducted. Using an i-vector Automatic Speaker Recognition (ASR) system based on perceptually relevant phonetic features, a recent study by Gerlach et al. (2020) explored the relationship between listener ratings of voice similarity and automatically obtained voice similarity estimates using scores output by the ASR system. Their stimuli consisted of paired speech samples of speakers from *DyViS*, an accent-controlled database containing Standard Southern British English (SSBE) male speech (Nolan et al. 2009). 106 listeners (mainly L1 speakers of British English or German) submitted voice similarity ratings of pairings of ten speakers online using a Likert scale. The study's results showed a broadly linear, statistically significant positive correlation between these listeners' ratings and the ASR similarity scores for the small group of speakers.

The present study expands on the results from Gerlach et al. (2020) by investigating further correlations using six additional speaker groups from three English-language databases (*DyViS*, *YorViS* (McDougall et al. 2015), *WYRED* (Gold et al. 2018)), which all contain speech from male speakers aged 18-30. The six speaker groups (referred to as *DyViS* D1/D2/D3, *YorViS* Y, and *WYRED* W1/W2) each consisted of 15 speakers. Each listener was asked to rate the similarity of all pairings of the 15 voices (including same-speaker pairs) in one of the six speaker groups on a 9-point Likert scale (120 comparisons), using samples of spontaneous speech, about 3s in duration. 20 listeners (aged 18-40, balanced for sex, English L1 speakers) were allocated to each speaker group.

The ASR comparison scores were obtained using five VOCALISE speaker modelling approaches (i.e. sessions): i-vectors based on auto-phonetic (automatically extracted phonetic) features, with and without F0, i-vectors based on spectral (MFCC) features, x-vectors based on auto-phonetic features (without F0), and x-vectors based on spectral features (see Alexander et al. 2016; Kelly et al. 2019). Stimuli for the ASR comparisons contained spontaneous speech and had a duration of approximately 4 minutes. None of the speaker databases used in the experiment were used to train the VOCALISE sessions.

The present study thus explores whether the correlation between voice similarity ratings and automatically obtained estimates found in Gerlach et al. (2020) holds across different same-accent speaker groups as well as different-accented speaker groups with listener ratings from more controlled listener groups. Additionally, it compares results from different ASR system speaker modelling approaches.

First results for speaker group D1 (Figure 1) show a stronger statistically significant correlation for the different-speaker (DS) comparisons than shown in Gerlach et al. (2020),

using the same VOCALISE session (i-vectors, auto-phonetic features with F0): Spearman's *rho* (2-tailed) = 0.484, $p < 0.001$, and linear regression resulted in RSQ = 0.253 ($p < 0.001$). Correlations for the five further speaker groups are also positive and significant for the majority of speaker modelling approaches; results using the x-vectors based on auto-phonetic features being particularly promising. This finding shows that the positive correlations hold across speaker groups with the same and with different accents, and provides further support for the use of automatic systems like VOCALISE, which take into account perceptually relevant features, in the assessment of speaker similarity.

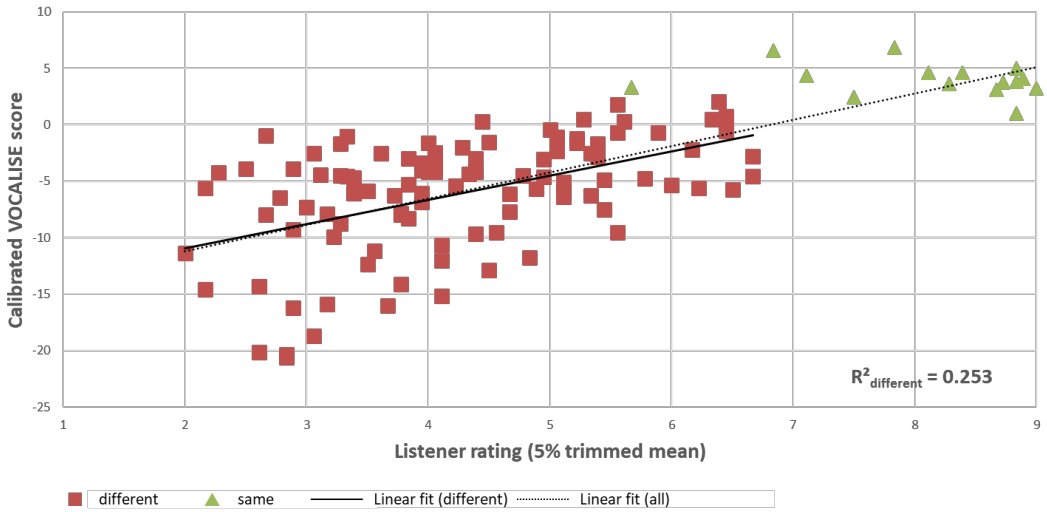

Figure 1 Scatterplot displaying the relationship between calibrated VOCALISE scores and the 5% trimmed mean of the listener ratings of speaker group *DyViS* D1 (Listener rating 1 = very dissimilar, 9 = very similar; DS comparisons in squares and same-speaker comparisons in triangles).

References

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, Bilbao, Spain.*

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.

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. Interspeech 2018, September 2-6 2018, Hyderabad*. 2748-2752.

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In *Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.*

McDougall, K. (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *IJSLL* 20(2): 163-172.

McDougall, K., Duckworth, M., & Hudson, T. (2015). Individual and group variation in disfluency features: a cross-accent investigation. In *Proc. 18th ICPhS, Glasgow.*1-5.

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *IJSLL* 16(1): 31-57.