Automatic assessment of voice similarity within and across speaker groups with different accents

Linda Gerlach[1,2], Kirsty McDougall[1], Finnian Kelly[2], Anil Alexander[2]

[1]University of Cambridge, United Kingdom, [2]Oxford Wave Research, United Kingdom
{lg589|kem37}@cam.ac.uk, {finnian|anil}@oxfordwaveresearch.com

ABSTRACT

As larger speaker databases materialise and speech technology becomes more efficient, an opportunity arises to automate assessments of perceived voice similarity ('PVS') among selections of voices for applications such as voice synthesis and forensic voice parades. Expanding on previous research [1], the present study addresses whether the correlation observed between listener ratings and automatic estimates of PVS can be found within different speaker groups with the same accent and across speaker groups with a different accent of English. Further, the impact on this correlation of variations in the automatic approach for PVS, based on a pre-trained automatic speaker recognition system, is explored; specifically, combinations of different feature extraction methods, speaker modelling approaches, and distance measures are considered. Results are positive and statistically highly significant within and across speaker groups particularly when relying on automatically extracted perceptually relevant phonetic features, demonstrating the generalisability of the method.

Keywords: voice similarity, voice parades, forensic speech science, automatic speaker recognition

1. INTRODUCTION

The assessment of perceived voice similarity ('PVS') between speech samples is a costly and time-consuming task when conducted manually. Subjective ratings from a number of listeners are required to approximate a 'ground-truth' value of PVS for a speaker pair, e.g. [2].

For forensic voice parades, where a witness is asked whether they can identify a perpetrator's voice among a number of foil voices that sound similar with regard to their pitch, speed, accent and more, assessing PVS is crucial for ensuring a fair selection of foils [3]. Under the scrutiny of an expert phonetician, partial automation of the selection process could allow for a larger pool of potential foils to be considered while facilitating a more objective selection at greater speed.

Voice banking applications aim to provide a person who has lost their voice, e.g., due to a medical condition, with a personalised speech synthesis device [4]. If audio recordings from the patient are not sufficient or unavailable, a 'donor voice' may be used for synthesis. Automatic assessment of PVS between the patient's voice (recordings) and a pool of donor voices may help swiftly narrow down suitable candidates for the patient to choose from and find their ideal donor voice.

Recent developments in voice synthesis show a need for fast assessments of similarity for large numbers of target speakers and their synthesised samples. Das et al. [5] explored the use of an automatic speaker recognition (ASR) system based on x-vectors for similarity assessments with encouraging results. In a similar vein, Deja et al. [6], trained a model based on spectral features to automatically evaluate PVS and yielded Pearson correlations of up to 0.78 between their model's output and perceptual ratings.

Past research also considered the use of phonetic features in automatic assessments of PVS [1, 7], but used only a small number of speakers. Building on Gerlach et al. [1], the present paper proposes assessing PVS using an existing pre-trained ASR system. This paper will address 1) whether there is a correlation between listener ratings and automatic estimates of PVS within different groups of speakers with the same accent and 2) whether this correlation can also be found across groups of speakers for different accents of English. It will further be explored 3) how the correlations observed are impacted by different feature extraction and speaker modelling approaches, as well as distance measures.

2. METHOD

2.1. Speaker databases and groups

Six speaker groups were chosen from DyViS [6, 7], YorViS [10], and WYRED [11]. DyViS contains 100 male speakers of Standard Southern British English (SSBE) aged 18-25. Three sets of 15 DyViS speakers (henceforth D1, D2, and D3) were employed to enable observation of variability between speaker groups of the same accent. Further, 15 male speakers of York English from YorViS (referred to as Y), also 18-25, were used. Finally, two groups of 15 speakers (hereafter W1 (Bradford) and W2 (Wakefield)) out of

180 male speakers (18-30 y.o.) of West Yorkshire Englishes were selected from WYRED. This selection enables the analysis of variation between three Yorkshire accent groups as well as comparison with the three same-accent SSBE groups. The recordings from all three databases contained spontaneous speech in studio quality from a mock police interview and a telephone call with an 'accomplice'.

2.2. Listener experiment

Listener ratings of PVS for all six speaker groups, collected as part of the VoiceSim [12], YorViS [10], and IVIP [13] projects, and presented in [14], were made available for the present study. For the listener experiments, the researchers had created two short voice samples of about 3s from each speaker's telephone call recording (studio quality). The samples were paired to form 120 speaker pairs (including same-speaker) for each speaker group. Listeners were then asked to rate the voice similarity of each speaker pair on a 9-point scale (1 – very similar, 9 – very different). Each speaker group was rated by a different set of 20 listeners (aged 18-40, English L1, born and raised mostly in England until aged 18), reporting no hearing impairments, and approximately balanced for sex.

For the present study, the given listener ratings were inverted (1 – low similarity, 9 – high similarity) to ensure that the scales used for the automatic estimation of voice similarity and the listener ratings were in the same direction.

2.3. Automatic comparisons

VOCALISE forensic ASR software [13, 14] was used in this experiment to obtain similarity scores automatically. For speaker modelling, the software provides pre-trained i-vector and x-vector frameworks that can use spectral Mel-frequency cepstral coefficients (MFCCs) or so-called auto-phonetic (AP, automatically extracted phonetic) features, which are regarded as perceptually relevant [7]. Independent of the feature extraction and speaker modelling approaches, in the automatic speaker comparison two speaker models are compared and a comparison score is calculated. In this study, the comparison scores are interpreted as indicators of voice similarity (low score – low similarity, high score – high similarity).

VOCALISE provides combinations of pre-trained modelling approaches and parameters as 'sessions' [15]. Table 1 shows the VOCALISE sessions tested in this study. I-vector session (a) with its combination of auto-phonetic features, including F0 among others, was included as this had previously been applied in similar experiments by [7] and [1] and is considered to be the most relevant session with regard to PVS (see also [16] regarding the importance of F0 and formant frequencies in PVS, and [12]). I-vector session (b) was included to evaluate the relevance of formant frequencies and F0 for the assessment of voice similarity. The i-vector session (c) based on MFCCs was included as similar approaches had been used previously, for instance in [18], and to test the hypothesis that approaches incorporating perceptually relevant phonetic features are superior to those that do not. Due to the development of new speaker modelling approaches, this experiment also includes x-vector sessions based on Deep Neural Networks. X-vector session (d) using spectral features outperforms i-vectors in speaker recognition tasks [16] and it is hypothesised that this speaker modelling approach may also have an impact on the representation of PVS. An x-vector session relying on auto-phonetic features (e) is explored as the inclusion of perceptually relevant features yielded promising results in previous experiments assessing voice similarity [1, 15]. Finally, an x-vector session using auto-phonetic features including F0 (f) is explored, and results compared to its i-vector equivalent (a).

	Feature extraction	Speaker model
a	Auto-phonetic (AP), ie. F0, semi-tones of F0, derivatives, LTF1 to LTF4	i-vector
b	AP, ie. LTF1 to LTF4	i-vector
c	spectral/MFCC	i-vector
d	spectral/MFCC	x-vector
e	AP, ie. LTF1 to LTF4	x-vector
f	AP, ie. F0, semitones of F0, derivatives, LTF1 to LTF4	x-vector

Table 1: Feature extraction and speaker modelling combinations used in VOCALISE.

Comparison scores were initially calculated based on the widely used PLDA (probabilistic linear discriminant analysis) which has great discriminatory power to tell speakers apart [16]. For comparison, cosine distance was also evaluated for calculating similarity scores to explore whether a less powerful discriminator may be more useful for assessing voice similarity.

Comparisons were conducted between all speakers within each of the six speaker groups. For each speaker, two samples (~4min) were used, one each from the telephone call and the interview task. The resulting comparison scores were then interpreted as voice similarity estimates. The scores were calibrated using Bio-Metrics [19] in order to normalise their numerical range and ensure comparability across speaker groups. The two

comparison scores for each speaker pair (spk1_1 vs spk2_2, spk2_1 vs spk1_2) were averaged to obtain one score per speaker pair [1].

2.4. Evaluation and analysis

To adjust for bias, the 5% trimmed mean of the listener ratings was calculated. Spearman rank correlation was performed to explore the degree and direction of the relationship between listener ratings of PVS and automatically obtained similarity estimates. To assess the potential for establishing thresholds for PVS in the future, the linearity of the relationship was measured using Pearson's correlation coefficient.

3. RESULTS

In sections 3.1. and 3.2. below, automatic comparison results are based on the i-vector session using auto-phonetic features including F0 as described in section 2.3. (a), and PLDA. Section 3.3. gives an overview of the performance of all VOCALISE sessions using PLDA to calculate scores, while section 3.4. considers the impact of PLDA versus cosine distance.

3.1. Same-accent background

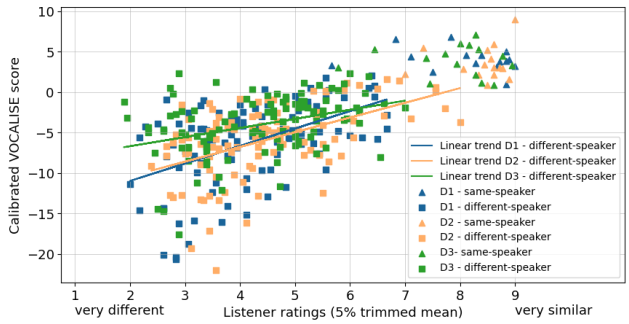

Figure 1: Scatter plot of listener ratings versus automatically obtained scores based on session (a) as described in 2.3. for DyViS D1 (blue), D2 (yellow), and D3 (green). DS comparisons are displayed as squares, SS comparisons as triangles.

Figure 1 illustrates the results for the three DyViS speaker groups, with listener ratings on the x-axis, and automatically obtained similarity estimates on the y-axis. Linear trendlines are shown for the different-speaker (DS) comparisons only. The graph shows that listeners seem to treat same-speaker (SS) comparisons in a similar way to DS comparisons, as listeners used a range of scores and did not rate SS comparisons consistently as 'very similar' across the three speaker groups. The automatic system showed overall good discrimination between SS and DS comparisons. The degree of the linear relationship between automatic scores and listener ratings varies across the three SSBE speaker groups, nevertheless, the arising trends all show a positive, largely linear relationship. In the statistical analyses which follow only DS comparisons are included.

3.2. Different-accent background

Speaker group	Spearman's rho	Pearson's r
D1	0.484**	0.503**
D2	0.519**	0.502**
D3	0.357**	0.366**
Y	0.463**	0.704**
W1	0.380**	0.398**
W2	0.552**	0.631**

Table 2: Correlation coefficients for each speaker group using the AP i-vector session with F0 and PLDA to approximate listener PVS ratings. Level of statistical significance highlighted as **<0.001, 2-tailed.

Table 2 shows that across all six speaker groups, the correlations between listener ratings and automatically obtained scores are statistically highly significant. For Spearman analyses, the lowest correlation is observed for D3, whereas W2 yielded the highest correlation, followed by D2. These results show that there is some variation between different speaker groups regardless of accent. This is also true for variation in linearity as indicated by the Pearson results; note that for Y, a very distinctive speaker had a substantial influence on the correlation.

3.3. Session performance across speaker groups

	Session	Spearman's rho	Pearson's r
a	i-vector AP inc. F0	0.457**	0.472**
b	i-vector AP	0.276**	0.269**
c	i-vector MFCC	0.349**	0.339**
d	x-vector MFCC	0.196**	0.201**
e	x-vector AP	0.417**	0.399**
f	x-vector AP inc. F0	0.244**	0.272**

Table 3: Correlation coefficients for each VOCALISE session based on automatic similarity estimates using PLDA and listener ratings of five speaker groups (D1, D2, D3, W1, W2). Level of statistical significance highlighted as **<0.001, 2-tailed.

To evaluate the performance of individual VOCALISE sessions, DS scores from all speaker groups except YorViS were combined and correlated with the aggregate of the corresponding 5% trimmed mean listener ratings. The Y speaker group was excluded as it was not possible to obtain calibrated scores for this set.

Table 3 shows that correlations for all sessions tested were statistically highly significant and broadly linear. X-vector session (d) relying on spectral features yielded the lowest Spearman correlation coefficient, followed by the i-vector session using auto-phonetic features without F0 (b). The next highest correlation coefficient was found using the i-vector session with spectral features (c). The sessions performing best in approximating ratings of PVS were the x-vector session using auto-phonetic features (e) and, ultimately, the i-vector session relying on auto-phonetic features and F0 (a). Of further note is that for the two best performing sessions, Spearman correlations for all individual speaker groups were positive and statistically highly significant, while this was not the case for the other sessions.

Based on these results, it was expected that a session using x-vectors and auto-phonetic features including also F0, semitones of F0 and derivatives (f) would have the potential to outperform the i-vector session using these features. Hence, a session combining these features for an x-vector model was also evaluated. In fact, results showed a statistically highly significant but low correlation between listener ratings of PVS and the new calculated scores compared to the majority of the previously tested sessions.

3.4. Impact of distance measure on correlation

Since PLDA as well as cosine distance are commonly encountered distance measures in speaker recognition, the correlation experiment using the five speaker groups (excl. Y) was repeated using cosine distance instead of PLDA.

Overall, again all correlation coefficients were positive and statistically highly significant. On average, Spearman correlation coefficients were higher using PLDA. On an individual session level, using cosine distance had an adverse effect on the ranking of the i-vector session using auto-phonetic features including F0, leaving it in third place (rho=0.309, $p<0.001$, 2-tailed), while the i-vector session using spectral features came up second best (rho=0.318, $p<0.001$, 2-tailed). Results showed a slight increase of 0.023 for the correlation coefficient for the x-vector session using auto-phonetic features (rho=0.440, $p<0.001$, 2-tailed), suggesting it as the best choice when using cosine distance.

4. DISCUSSION

This study expanded on [1] to show the generalisability of the proposed voice similarity assessment across more diverse speaker groups. Overall, the relationship between listener ratings and automatic estimates of PVS was found to be positive and statistically significant within the speaker groups of the same accents as well as across speaker groups of a different accent of English. Variability was present in the strength and linearity of the correlations, but there were no apparent trends related to the speakers' accents, likely due to the variation of individual speakers or listeners exceeding any variation related to the different accents and impact of sample choice.

The best performance in automatically approximating listener ratings of PVS using a pre-trained ASR system was yielded by applying the i-vector session incorporating LTF1 to LTF4, F0 and semitones, as well as derivatives in combination with PLDA as a distance measure, closely followed by the x-vector session drawing solely on LTF1 to LTF4 in combination with PLDA. The latter session's correlation with the listener ratings improved further when using cosine distance. Further conclusions can be drawn if interrater agreements can be better assessed.

While the similarity rating scales used by Deja et al. [6] (0-100) and in the current experiment (1-9) may not be directly comparable, the correlations in this study are not as high as those reported by Deja et al.. Additional experiments considering synthetic speech are needed to draw further conclusions.

Regarding the different applications mentioned in section 1., using an existing ASR system with features that are relevant for PVS may offer an opportunity to select similar-sounding natural voices for voice parades and voice banking. It should be noted that the search space for similar-sounding voices may need to be restricted in terms of demographic data to be appropriate for the task. Further, a suitable degree of similarity between voices acceptable for a particular application must be quantified, e.g., by using calibrated scores from the automatic system.

5. CONCLUSION

This paper showed that using a pre-trained ASR system to assess PVS provides positive, statistically highly significant results within and across speaker groups of different accents of English, particularly using automatically extracted, perceptually relevant phonetic features. Future work will investigate score thresholds for PVS for different applications and explore large and diverse speaker databases using similarity scores and clustering methods.

6. ACKNOWLEDGEMENTS

The authors thank Oscar Forth for his assistance with VOCALISE modifications for the paper. This work was partially funded by the Selwyn Cambridge – Oxford Wave Research PhD Studentship in Forensic Phonetics and Automatic Speaker Recognition and the IVIP project 'Improving Voice Identification Procedures' which is funded by the UK Economic and Social Research Council, reference ES/S015965/1.

7. REFERENCES

[1] 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 Commun.*, 124, 85–95. doi: 10.1016/j.specom.2020.08.003.

[2] McDougall, K., 2013. Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. *Int. J. Speech Lang. Law*, 20(2), 163–172. doi: 10.1558/ijsll.v20i2.163.

[3] Home Office, 2003. Advice on the use of voice identification parades. [Online]. Available: https://tinyurl.com/26r4ucn6 (28.04.2023).

[4] Yamagishi, J., Veaux, C., King, S., Renals, S., 2012. Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction. *Acoust. Sci. Technol.*, 33(1), 1–5. doi: 10.1250/ast.33.1.

[5] Das, R. K., Kinnunen, T., Huang, W., Ling, Z., Yamagishi, J., Zhao, Y., Tian, X., Toda, T., 2020. Predictions of subjective ratings and spoofing assessments of Voice Conversion Challenge 2020 submissions. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 99–120. doi: 10.21437/VCC_BC.2020-15.

[6] Deja, K., Sanchez, A., Roth, J., Cotescu, M., Jul. 2022. Automatic evaluation of speaker similarity. In *Proc. of INTERSPEECH 2022* Incheon, Paper 75. doi: 10.48550/arxiv.2207.00344.

[7] 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. In *Proc. of INTERSPEECH 2016* San Francisco, Paper 2018, 1567–1568. Available: https://tinyurl.com/mr3nyhhr (28.04.2023).

[8] Nolan, F., Dynamic Variability in Speech: a forensic phonetic study of British English, 2006-2007. 2011, UK Data Service. SN: 6790. doi: http://doi.org/10.5255/UKDA-SN-6790-1.

[9] Nolan, F., McDougall, K., de Jong, G., Hudson, T., 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Int. J. Speech, Lang. Law*, 16(1), 31–57. doi: 10.1558/ijsll.v16i1.31.

[10] McDougall, K., Duckworth, M., Hudson, T., 2015. Individual and group variation in disfluency features: a cross-accent investigation. *Proc. 18th ICPhS*, Paper 0308. Available: https://tinyurl.com/jw2v3uu9 (28.04.2023)

[11] 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. In *Proc. of INTERSPEECH 2018* Hyderabad, Paper 0065, 2748–2752. doi: 10.21437/Interspeech.2018-65.

[12] Nolan, F., McDougall, K., Hudson, T., 2011. Some acoustic correlates of perceived (dis)similarity between same-accent voices. In *Proc. 17th ICPhS* Hong Kong, 1506–1509. Available: https://tinyurl.com/yf9skafm (28.04.2023).

[13] Pautz, N., McDougall, K., Mueller-Johnson, K., Nolan, F., Paver, A., Smith, H. M. J., 2023. Identifying unfamiliar voices: examining the system variables of sample duration and parade size. *Q. J. Exp. Psychol.*, 1–19. doi: 10.1177/17470218231155738.

[14] McDougall, K., 2021. Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence. In *Proc. of XVII AISV (Associazione Italiana Scienze della Voce) conference: "Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications"*, 33–56. doi: 10.17469/O2108AISV000002.

[15] Alexander, A., Forth, O., Atreya, A. A., Kelly, F., 2016. VOCALISE : a forensic automatic speaker recognition system supporting spectral , phonetic , and user-provided features. In *Odyssey 2016 Show & Tell* Bilbao, Paper 88. Available: https://odyssey2016.org/papers/Show_tell/88.pdf (28.04.2023)

[16] 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 *Proc. of the AES International Conference* Porto, Paper 27. Available: http://www.aes.org/e-lib/browse.cfm?elib=20477 (28.04.2023).

[17] McDougall, K., 2013. Earwitness evidence and the question of voice similarity. *Br. Acad. Rev.*, 21, 18–21. Available: https://tinyurl.com/47umjb3u (28.04.2023).

[18] Lindh, J., Eriksson, A., 2010. Voice similarity - a comparison between judgements by human listeners and automatic voice comparison. In *Proc. of FONETIK 2010* Lund, 63–68. Available: https://journals.lub.lu.se/LWPL/article/view/19647/17775 (28.04.2023)

[19] Bio-Metrics 1.8 Performance Metrics Software. 2019, Oxford Wave Research. Available: https://oxfordwaveresearch.com/products/bio-metrics/ (28.04.2023).