INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Voice twins: discovering extremely similar-sounding, unrelated speakers

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

This paper deals with extremely similar-sounding, unrelated speakers ('voice twins') and presents an automatic approach to voice twin discovery applied to different speaker databases. An automatic speaker recognition system relying on perceptually relevant phonetic features including formants and a tuned clustering algorithm DBSCAN was used to group recordings within diverse datasets. 18 voice twin pairs selected from 2-speaker clusters were evaluated by 50 listeners in a 2-alternative forced choice experiment. Same/different decisions and confidence ratings were collected for same-speaker, random different-speaker and voice twin comparisons. Listeners were unable to differentiate between the candidate voice twin pairs much better than chance level while they performed well (80% accuracy) for random same- or different-speaker comparisons indicating the voice twin speakers were perceptually very similar. The implications and forensic relevance of identifying voice twins are discussed.

Index Terms: forensic phonetics, voice similarity, voice perception, voice twins, speaker clustering

1. Introduction

While twins and speakers related to one another have been the focus of many studies exploring how to distinguish similar speakers e.g., [1, 2], little effort has gone into systematically locating extremely similar-sounding but unrelated speakers (henceforth referred to as 'voice twins'). Finding voice twins could offer further insights into perceived voice similarity and speaker individuality. Additionally, a systematic approach to discovering voice twins within a larger group of speakers could be applied to select the most alike donor voice from a voice bank for a person with a voice or speech impairment, as well as to assess how good a synthesised voice is with regard to a target speaker. In a forensic domain, voice twins may be utilised as difficult speaker comparisons in ear-witness assessment tasks [3] or in the quest to find super-recognisers [4].

In a database containing many speakers there is a greater likelihood of finding extremely similar-sounding speakers. An automatic approach to assess voice similarity would allow such similar-sounding speakers to be identified easily. One such approach was recently proposed by Deja et al. [5] focussing on the comparison of synthetic and natural speech and gave promising results. In forensic phonetics, Fröhlich et al. [6] used a pre-trained ECAPA-TDNN model in combination with F0 deltas to assess the similarity of speaker pairs, while Schäfer and Foulkes [3] focused mainly on long-term F0 measures. In these studies, discrimination tasks by listeners revealed a 13.8% difference in accuracy between random different-speaker comparisons and difficult different-speaker comparisons [6], and showed differences in listener abilities [3]. Gerlach et al. [7] evaluated a pre-trained automatic speaker recognition system based on perceptually relevant phonetic features to approximate listener ratings of perceived voice similarity and achieved positive, highly significant correlations.

In the present study, the approach by Gerlach et al. [7] is expanded on by combining the automatic speaker recognition system with a highly tuned clustering approach to explore whether voice twins, i.e. unrelated, extremely similar-sounding speakers, indeed exist. It investigates how well listeners can distinguish between voice pairs comprising automatically selected voice twins (VT comparisons) compared with randomly chosen different-speaker (DS) comparisons and same-speaker (SS) comparisons, and whether listeners' confidence in their decisions is affected by the type of comparison they face. Further, it assesses whether the perceived similarity within speaker pairs varies across the three comparison types (VT, DS, SS).

2. Method

2.1. Speaker databases

Experimental stimuli were developed using three speaker databases of different sizes and variability in terms of speaker demographics and recording conditions. WYRED [8] is a highly controlled database containing 180 male English speakers aged 18-30 from Bradford, Wakefield, and Kirklees in West Yorkshire, England. A subset of studio quality recordings with spontaneous speech was used (Tasks 2 and 4, each containing one file per speaker).

The GBRENG database [9] contains 6000 landline and mobile telephone recordings of spontaneous speech from 600 male and female adult speakers. Two subsets of good quality landline recordings of speakers brought up in the UK were selected (net speech >30 s, WADA SNR [10] >24 dB). The female subset contained 1,165 recordings of 208 speakers, the male one 1,067 recordings of 193 speakers.

VoxCeleb1 [11, 12] is a large and diverse database of celebrity recordings in varying conditions and background noise. The database is shared as YouTube URLs; 10,152 recordings from 1,250 speakers were gathered. Good quality recordings (net speech >20 s, WADA SNR >18 dB) were selected to make up a set of 505 female speakers (1,796 recordings) and a set of 609 male speakers (2,242 recordings). Lower audio quality thresholds were chosen for VoxCeleb1 than for GBRENG to allow for higher variability. In total, across three databases, 6,576 recordings from 1,695 male and female speakers were processed in this study.

10.21437/Interspeech.2023-2134

2.2. Clustering experiment

Each of the five database subsets was subjected to the following procedure: First, perceptually relevant phonetic features (long-term formant (LTF) distributions of F1 to F4) were extracted. Next, the features were passed to a trained x-vector DNN [13], and an x-vector was obtained for each recording. Similarity scores were calculated based on cosine similarity between all recordings within a subset using VOCALISE automatic speaker recognition software [13]. A tuned DBSCAN algorithm (density-based spatial clustering of applications with noise [14]) was used to cluster x-vectors based on their proximity to each other within areas of high density, determined by a set threshold of neighbouring points that fall within a specific radius (epsilon, henceforth ϵ). The ϵ was varied from 0.1 to 0.3 based on the observation that below 0.1 even same-speaker clusters barely emerged and above 0.3 clusters formed containing numerous speakers unlikely to sound very similar.

For this experiment, voice twin candidates were taken from clusters containing two speakers with two or more recordings each. Clusters with more than two speakers were not considered for the sake of simplicity. Table 1 displays the number of potential voice twin clusters found within each database subset at each ϵ value tested.

Table 1: *Number of candidate voice twin clusters per database subset (row) and ϵ (column).*

	0.10	**0.15**	**0.20**	**0.25**	**0.30**	**Total**
WYRED	0	0	1	4	4	9
GBRENG - M	0	0	2	6	9	17
GBRENG - F	0	1	2	3	2	8
VoxCeleb - M	0	0	4	3	0	7
VoxCeleb - F	0	1	0	1	1	3

In order to select the most similar-sounding voice twins, the voice twin candidates obtained using the lowest ϵ possible within each database subset were chosen. As only three pairs of voice twins were found for the female VoxCeleb subset, the other subsets were limited to a selection of three pairs of voice twins to balance out the number of samples. If there were more than three pairs of voice twins with an equally low ϵ available, three pairs were chosen randomly. To reduce the effects of language differences and to focus mainly on the voice itself, if language varieties differed within a voice twin pair (e.g. New Zealand vs British English), a new voice twin pair was randomly selected. Where there were more than two files per speaker, the files with the overall highest automatic similarity scores were selected to create samples for the listener experiment. No speaker overlap was observed between voice twin candidates at different ϵ.

2.3. Listener experiment

For each speaker within a voice twin pair, from the same database subset, a recording of a different speaker with the same language variety but not a member of any voice twin pairs was chosen at random and a 3s-sample manually created. Four 3s-samples were also taken from each speaker (two per file) within the voice twin pairs and three of these files were randomly selected to make up six comparisons per voice twin pair: two voice twin comparisons, two DS comparisons, and two SS comparisons. For each database subset, 18 comparisons were assembled (90 comparisons overall) for a two-alternative forced choice (2AFC) task. Samples were controlled to contain normal speech (i.e. no laughter) and as little background noise as possible. Samples were amplitude-normalised and padded with 100ms of silence at the beginning and end.

The experiment was conducted online using Gorilla [15]. The experiment contained a consent form and a short metadata questionnaire collecting details on age, sex, language background and linguistic knowledge. A headphone screening [16] was included before the start of the comparison task. Listeners were able to familiarise themselves with the task based on five comparison pairs that reflected the databases and recording conditions encountered in the experiment. The actual test phase consisted of five blocks of 18 comparisons each, separated by short breaks, in the order of WYRED - 2x GBRENG - 2x VoxCeleb. While the blocks could contain male and female speakers from the respective databases, the order within the blocks was randomised. Further, listeners were randomly assigned to one of two sample orders. In each trial consisting of two consecutively played samples, listeners were asked to indicate whether they thought the voices came from the same or different speakers and how confident they were in their decision on a 5-point Likert scale (1 = not confident at all, 5 = very confident). The experiment did not allow listeners to replay the samples. At the end of the experiment, listeners were given feedback on their performance and asked whether they recognised any speakers.

2.4. Participants

Ethics approval was obtained from the Faculty of Modern and Medieval Languages and Linguistics at the University of Cambridge. Fifty participants (25 male, 25 female) were recruited and paid for their time using Prolific [17]. The participant pool was limited to those aged 18 to 40 years, with English as their first language, without hearing impairments, and who were born in the UK and spent most of their time before turning 18 there. Participants with high approval ratings on the platform and an adequate internet connection speed in Chrome browser were used.

2.5. Evaluation and analysis

To assess how similar the automatically selected voice twin candidates were, firstly listeners' discrimination performance across the three comparison types (DS, SS, VT) was examined and differences between comparison types were assessed using a Welch's ANOVA. As listeners' confidence in their discrimination abilities may vary based on comparison type, particularly since they will not be as familiar with comparing similar-sounding voices from different speakers, confidence ratings were analysed. A Kruskal-Wallis H test was used to check whether there were relevant differences in the distributions of confidence ratings between any of the three comparison types. Dunn post hoc test with Holm correction was applied to confirm which specific groups differed significantly and adjusted p-values are reported. To assess confidence in combination with listeners' same/different decisions, similar to the method used by Afshan et al. [18], a score of perceived similarity was calculated for each comparison. Confidence ratings were recoded as 0 to 4 and then multiplied by the corresponding same/different response (-1 for 'different', 1 for 'same') resulting in a continuous scale from -4 (high level of confidence that the voices were from different speakers) to 4 (high level of confidence that the voices were from the same speaker) with a midpoint of 0 denoting a complete guess/no decision. Differences between groups were again assessed with a Kruskal-Wallis H and Dunn post hoc test. Statistical analyses were performed using R statistical software (v.4.1.2) [19].

3. Results

3.1. Initial checks

Participants used the points of the rating scale with varying frequency to indicate their confidence during the same/different decision task and engaged sincerely with the experiment in that they did not click only 'SAME' or only 'DIFFERENT' throughout the task. Overall listener agreement for same/different decisions as per Krippendorff's alpha for nominal data where 0 is complete disagreement and 1 is complete agreement resulted in $\alpha = 0.414$. This indicates rather high variability and could point at large differences between listeners regarding their speaker discrimination abilities.

3.2. Same/different decisions

3.2.1. All comparisons

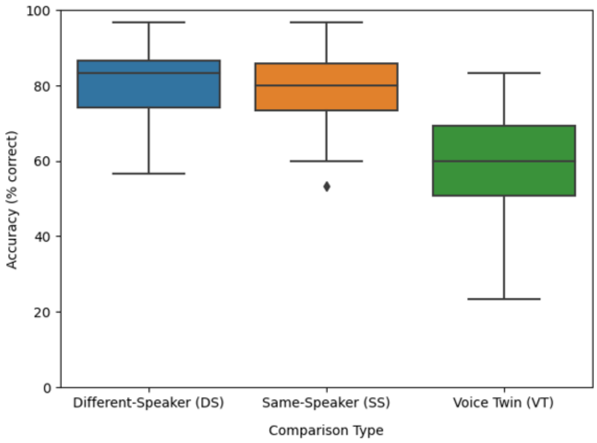

Figure 1: *Boxplots for listener accuracy (% correct) per comparison type.*

For all comparisons together, the mean accuracy (% correct) is 72.58% ($SD = 14.87\%$), however, there are differences between the comparison types DS, SS, and VT, as Figure 1 illustrates. Listeners performed similarly when facing DS ($M_{correct} = 80.27\%$, $SD_{correct} = 9.54\%$) and SS comparisons ($M_{correct} = 78.73\%$, $SD_{correct} = 9.85\%$). No listener correctly identified all DS comparisons as 'different' or all SS comparisons as 'same'. Regarding VT comparisons, listeners committed more errors, wrongly identifying more pairs as 'same' ($M_{correct} = 58.73\%$) and a greater standard deviation is apparent ($SD_{correct} = 13.78\%$). While no listener was deceived by all voice twin pairs, the maximum accuracy was lower in VT comparisons than in the other two comparison types. A Welch's ANOVA indicated significant differences between at least two groups. Games-Howell post hoc testing confirmed highly significant differences between DS and VT comparisons and between SS and VT comparisons ($p_{adjust} < .001$), but not between DS and SS comparisons ($p_{adjust} = .710$).

3.2.2. Individual differences between voice twin pairs

As voice twins were selected from multiple databases and based on different ϵ values, mean listener accuracy for the individual voice twins is shown in Figure 2. It can be seen that seven out of 15 voice twin pairs are wrongly recognised as the same speaker around or below chance level, including the two voice twin pairs with the lowest ϵ (0.15). The overall lowest accuracy of 24% is yielded with an ϵ of 0.15 by a female voice twin pair from VoxCeleb. Note that some candidate voice twin pairs, including the pair with the highest ϵ (0.3), were recognised correctly as different speakers with the comparatively high accuracy of 80-82%.

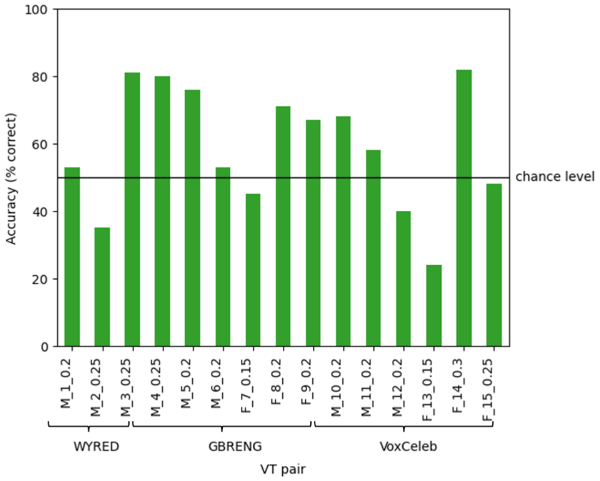

Figure 2: *Bar plot showing mean accuracy (% correct) for each voice twin pair. X-axis tick labels indicate sex followed by pair number and ϵ value; horizontal brackets refer to the respective database.*

3.3. Confidence ratings

Overall, participants were confident in their same/different decisions, but also used the full confidence rating scale from 1 to 5. Table 2 shows that listeners were slightly more confident when facing DS comparisons than SS comparisons. Listeners were least confident when facing VT comparisons. The distributions of confidence ratings in the three comparison types were found to differ significantly ($p_{adjust} < .001$).

Table 2: *Means, standard deviations, and medians of confidence ratings overall, and across the three groups DS, SS, and VT.*

	Mean	Standard deviation	Median
All groups	3.95	1.08	4
DS	4.14	1.03	4
SS	3.93	1.08	4
VT	3.76	1.09	4

3.4. Similarity scores

Violin plots (combining box plots and kernel density estimates (KDE)) displaying the distributions of similarity scores integrating same/different decisions and confidence ratings (see

Section 2.5) are given in Figure 3 for each comparison type. Across the three groups, it is apparent that listeners were confident independently of whether they made a same- or different-speaker decision. DS comparisons achieved a median similarity score of -3. A thin tail of the KDE indicates few DS comparisons received a high similarity score, i.e. highly confident same-speaker decisions. Regarding SS comparisons the distribution is inverted with a median similarity score of 3. VT comparisons show a different pattern as the KDE has a bimodal distribution, indicating that listeners gave ‘same’ and ‘different’ ratings with similar, comparatively lower confidence. With a median of -2, VT comparisons obtained a higher similarity score than DS comparisons. The differences between all groups were confirmed to be statistically highly significant ($p_{adjust} < .001$).

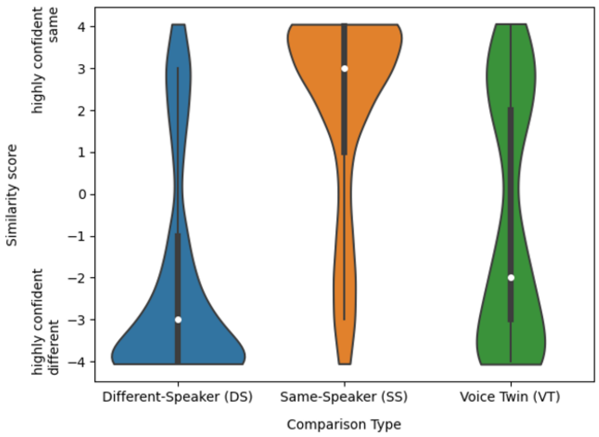

Figure 3: *Violin plot combining a boxplot and KDE indicating the distribution of similarity ratings for DS, SS, and VT comparisons. Median similarity ratings are signified by a white dot.*

4. Discussion

A listener experiment using automatically selected VT pairs and SS and DS control groups showed that significantly more detection errors were committed for the VT group, reflecting the difficulty posed by the similarity of the speakers. Overall accuracy roughly concurred with previous research using similar sample lengths e.g., [20]. The difference in mean accuracy between DS and VT comparisons ($\Delta_{accuracy} = 21.5\%$) was higher than that reported by Fröhlich et al. [6], indicating that the present method may produce more similar-sounding speaker pairs. A bigger standard deviation of error rates occurred in the VT group compared with DS and SS, indicating individual differences between listeners regarding their discrimination abilities when facing difficult speaker comparisons. This highlights an opportunity to locate ‘super-recognisers’. Overall, seven of 15 voice twin pairs were wrongly recognised as the same speaker around or above chance level; exploration of underlying features contributing to the difficulty of these comparisons remains for future work. In this study, one voice twin pair with the lowest ϵ yielded the lowest accuracy, while the voice twin pair with the highest ϵ was among the ones recognised with high accuracy. Confidence ratings were affected by comparison type, with listeners being least confident when confronted with VT comparisons, followed by SS comparisons. Listeners may rarely be exposed to voice twins in real life and are thus not trained in distinguishing between them. Differences in modelling intra- and inter-speaker variability with respect to voice twins require further consideration. Similarity ratings also differed significantly across comparison types, with listeners being less confident in their same/different decisions in VT comparisons, independent of the option they chose, while in DS and SS comparisons the incorrect option was accompanied by lower confidence. Further investigation into the relationship between confidence and correctness is required.

It is noted that there are several factors at play in automatically locating voice twins, including the recording conditions and sample lengths, number of speakers and recordings per speaker, and the distance metric used to evaluate similarity (e.g. ϵ). These factors may be varied depending on the requirements of the task being performed.

5. Conclusion

This study explored whether voice twins, i.e. unrelated, extremely similar-sounding speakers, exist, and offers an automatic approach for locating them. A listener experiment showed that while listeners performed well with 78-80% accuracy in same-speaker and random different-speaker comparisons, this was reduced to 58% in voice twin comparisons.

Listeners were also the least confident when assessing voice twin pairs compared to random different-speaker and same-speaker comparisons, independent of whether their decision was correct. These results demonstrate the successful selection of extremely similar-sounding speakers.

This research contributes to recent efforts towards creating difficult speaker comparisons to assess ear-witnesses or find super-recognisers and to developments towards automatically assessing speaker similarity for voice banking and voice synthesis applications.

6. Acknowledgements

This work was funded by the Selwyn Cambridge - Oxford Wave Research PhD Studentship in Forensic Phonetics and Automatic Speaker Recognition.

7. References

[1] H. J. Künzel, “Automatic speaker recognition of identical twins,” *International Journal of Speech, Language and the Law*, vol. 17, no. 2, pp. 251–277, 2010.

[2] E. San Segundo, A. Tsanas, and P. Gómez-Vilda, “Euclidean distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics,” *Forensic Science International*, vol. 270, pp. 25–38, Jan. 2017.

[3] S. Schäfer and P. Foulkes, “Assessing the individual voice recognition skills of earwitnesses,” in *Proceedings of British Association of Academic Phoneticians (BAAP) Colloquium*, York (online), UK, Apr. 2022.

[4] R. E. Jenkins, S. Tsermentseli, C. P. Monks, D. J. Robertson, S. V. Stevenage, A. E. Symons, and J. P. Davis, “Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks,” *Applied Cognitive Psychology*, vol. 35, no. 3, pp. 590–605, May 2021.

[5] K. Deja, A. Sanchez, J. Roth, and M. Cotescu, “Automatic evaluation of speaker similarity,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Incheon, South Korea, Sep. 2022, paper 75.

[6] A. Fröhlich, V. Dellwo, and M. Ramon, “Developing a challenging speaker discrimination test,” in *Proceedings of the VoiceID Conference*, Zürich, Switzerland, Jul. 2022.

[7] L. Gerlach, K. McDougall, F. Kelly, A. Alexander, and F. Nolan, "Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features," *Speech Communication*, vol. 124, pp. 85–95, 2020.

[8] E. Gold, S. Ross, and K. Earnshaw, "The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 2748–2752.

[9] GBR-ENG database, "A telephonic speech database collected for the UK government for evaluating speech technologies. Further details on application." 2019.

[10] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 2598–2601.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 2616–2620.

[12] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[13] F. Kelly, O. Forth, S. Kent, L. Gerlach, and A. Alexander, "Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors," in *Proceedings of the AES International Conference*, Porto, Portugal, 2019, paper 27.

[14] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, 2017.

[15] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior Research Methods*, vol. 52, pp. 388–407, Feb. 2020.

[16] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, and Psychophysics*, vol. 79, pp. 2064–2072, Oct. 2017.

[17] "Prolific," data collection 15.-22.12.2022. [Online]. Available: https://www.prolific.co/

[18] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech," *The Journal of the Acoustical Society of America*, vol. 151, pp. 1393–1403, 2022.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[20] A. Bartle and V. Dellwo, "Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech," *International Journal of Speech, Language and the Law*, vol. 22, pp. 229–248, Nov. 2015.