Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features.

Linda Gerlach[a], Philipps-Universität Marburg and Oxford Wave Research
Kirsty McDougall[b], University of Cambridge
Finnian Kelly[c], Oxford Wave Research
Anil Alexander[d], Oxford Wave Research
Francis Nolan[e], University of Cambridge

a Oxford Wave Research Ltd,
Clarendon Business Centre,
Sandford Gate, East Point Business Park, Sandy Lane West,
Oxford OX4 6LB
United Kingdom

linda@oxfordwaveresearch.com

b (Corresponding author)
Theoretical and Applied Linguistics Section
Faculty of Modern and Medieval Languages and Linguistics
University of Cambridge
Sidgwick Avenue
Cambridge CB3 9DA

kem37@cam.ac.uk

c Oxford Wave Research Ltd,
Clarendon Business Centre,
Sandford Gate, East Point Business Park, Sandy Lane West,
Oxford OX4 6LB
United Kingdom

finnian@oxfordwaveresearch.com

d Oxford Wave Research Ltd,
Clarendon Business Centre,
Sandford Gate, East Point Business Park, Sandy Lane West,
Oxford OX4 6LB
United Kingdom

anil@oxfordwaveresearch.com

e Theoretical and Applied Linguistics Section
Faculty of Modern and Medieval Languages and Linguistics
University of Cambridge
Sidgwick Avenue
Cambridge CB3 9DA

fjn1@cam.ac.uk

Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features.

Linda Gerlach, Philipps-Universität Marburg and Oxford Wave Research

Kirsty McDougall, University of Cambridge

Finnian Kelly, Oxford Wave Research

Anil Alexander, Oxford Wave Research

Francis Nolan, University of Cambridge

Abstract

The present study[1] investigates relationships between voice similarity ratings made by human listeners and comparison scores produced by an automatic speaker recognition system that includes phonetic, perceptually-relevant features in its modelling. The study analyses human voice similarity ratings of pairs of speech samples from unrelated speakers from an accent-controlled database (DyViS, Standard Southern British English) and the comparison scores from an i-vector-based automatic speaker recognition system using 'auto-phonetic' (automatically extracted phonetic) features. The voice similarity ratings were obtained from 106 listeners who each rated the voice similarity of pairings of ten speakers on a Likert scale via an online test. Correlation analysis and Multidimensional Scaling showed a positive relationship between listeners' judgements and the automatic comparison scores. A separate analysis of the subsets of listener responses from English and German native speaker groups showed that a positive relationship was present for both groups, but that the correlation was higher for the English listener group. This work has key implications for forensic phonetics through highlighting the potential to automate part of the process of selecting foil voices in voice parade construction for which the collection and processing of human judgements is currently needed. Further, establishing that it is possible to use automatic voice comparisons using phonetic features to select similar-sounding voices has important applications in 'voice casting' (finding voices that are similar to a given voice) and 'voice banking' (saving one's voice for future synthesis in case of an operation or degenerative disease).

Keywords

Perceived voice similarity, speaker similarity, automatic speaker recognition, voice parades, earwitness evidence

[1] Declaration of interest: The first, third and fourth authors, Linda Gerlach, Finnian Kelly and Anil Alexander, are employed by Oxford Wave Research Ltd. Oxford Wave Research develop and sell VOCALISE, the forensic automatic speaker recognition system evaluated in this submission.

1 Introduction

1.1 Perceived voice similarity

When listeners compare the voices of different speakers, they perceive certain speakers as sounding more similar to each other than others. The phenomenon of “perceived voice similarity” is not well understood in phonetic terms. The few studies available investigating the contributions of different phonetic features to the perception of voice similarity tend to offer limited findings due to the methodologies employed. For example, an early study by Walden et al. (1978) considered the perception of a single word only, spoken by different speakers. This experiment highlighted fundamental frequency and word duration as playing a role in voice similarity judgements. Work by Remez et al. (2007) on the read speech of a mixed-accent and mixed-sex group of speakers identifies dynamic formant frequency information as contributing to perceived voice similarity. In evaluating the results of a voice line-up experiment involving nine Gothenburg Swedish speakers, Lindh (2009) conducted a pairwise speaker similarity experiment which highlighted speaking tempo as a possible contributor to voice similarity judgements.

Perceptual similarity between the voices of Canadian French speakers is considered by Baumann and Belin (2010). Their study examines correlations between a range of acoustic features and speakers’ relative locations in a listener-determined perceptual space using Multidimensional Scaling (MDS), a data reduction technique which determines a number of pseudo-perceptual dimensions enabling the similarity of the objects (here, speakers) to be inferred (Schiffman et al., 1981). Baumann and Belin's study focusses on judgements of same/different speaker rather than judgements of the extent of similarity of different speakers and uses sustained vowels only. The authors reported highest correlations between perceptual measures and F0 and F1 (female speakers) and F0 and the mean difference between F4 and F5 (male speakers). A study which uses MDS analysis to analyse listener judgements on a similar-dissimilar Likert scale is that of Nolan et al. (2011); see also McDougall (2013). This study investigates the perceived similarity of a homogeneous group of male Standard Southern British English speakers and uses spontaneous speech stimuli. F0, vowel formants and a range of voice quality features are shown to be important factors in the judgement of voice similarity.

1.2 Automatically-determined speaker similarity

Stepping aside from human listeners, the extent of similarity between speakers can also be assessed by automatic speaker recognition (ASR) systems. ASR systems can be used to compare two speech samples and produce a score reflecting the likelihood that the speech samples were produced by the same speaker. While there are many approaches to ASR, it typically follows a sequence of feature extraction and speaker modelling, followed by the comparison of two speaker models, resulting in a numerical score. These numerical scores can be used to inform decisions about speaker identity (in commercial applications such as telephone banking, for example). In a forensic context, ASR systems can be used to calculate a likelihood ratio, which expresses the output score of the system under two competing hypotheses, e.g., that the speech samples originate from the same speaker, and that the speech samples originate from different speakers. The present

paper considers the use of ASR comparison scores not for decisions about identity or the calculation of likelihood ratios, but as a means of quantifying the similarity of different speakers. It addresses the question as to whether the assessments of speaker similarity made by an ASR system bear any correspondence to human listener judgements of voice similarity. If such relationships are present, automatically-determined voice descriptors should offer some insight into the acoustic workings of perceived voice similarity.

In one relevant study, Lindh and Eriksson (2010) automatically assessed the similarity of male speakers who had previously been used in the mock voice parade in Lindh (2009) and compared these assessments with those made by human listeners. Listeners were asked to indicate the dissimilarity of all pairings of nine voices on a Likert scale using an online experimental interface. The automatic comparison was conducted using a Gaussian Mixture Model-Universal Background Model (GMM-UBM) system. In addition to comparing the rank of each speaker when tested against another speaker model, the listener judgements and automatic results were compared using MDS. Lindh and Eriksson found some visual similarities between the two-dimensional plots they compared, yet further quantitative analysis was not explored. The authors noted that it was difficult on the basis of their results to draw clear conclusions about the extent of the relationship between the human and automatic assessments of similarity.

Previous work has also compared similarity ratings by humans and machines for the purpose of selecting voice actors for film dubbing, where a voice speaking a foreign language is inevitably compared with the voice of the original actor. Obin and Roebel (2016) compared two automatic approaches to finding voice actors from a set of 50 speakers. They used both an MFCC i-vector system and a system which calculated a score taking into account manually added labels. The labels provided information such as speaker age, gender, voice quality, and emotion. Listeners were asked to judge the similarity between an original voice and a set of voices selected by the automatic approaches. The results of the multi-label classification approach were closer to the listeners' performance than those of the i-vector system. However, the necessity of considerable (and time-consuming) manual labelling is a clear disadvantage of this approach.

An experiment using auto-phonetic features (automatically extracted F0 and formant frequencies, and their derivatives) in an i-vector based system to identify (dis)similar voices was conducted by Kelly et al. (2016). The study's stimuli were taken from SITW (Speakers in the Wild) lapel microphone recordings in a variety of English accents. However, the SITW database is not controlled for recording conditions and is biased towards male speakers (McLaren et al., 2016). For a group of three male and three female speakers, sets of similar and dissimilar speakers were obtained by an automatic comparison. In a web-based experiment, a group of 43 listeners judged the similarity of each speaker with its automatically assigned similar and dissimilar speakers, and with a different sample of the speaker itself. Male speakers and their similar comparison speakers received significantly higher similarity judgements from the listeners than male speakers and their dissimilar comparison speakers. The i-vector-based comparison scores and median listener similarity ratings were found to have a positive linear correlation for male speakers only. Results for female speakers were inconclusive. A major drawback of this study is the small sample size of speakers compared.

With the exception of Kelly et al. (2016), the few studies that have considered links between automatically-determined similarity and perceived voice similarity have used automatic approaches based on spectral features. It is nevertheless the case that the human perception of voice similarity relates to acoustic and voice quality features, as well as spectral features. The auto-phonetic approach used in Kelly et al. (2016) may be more comparable to human perception of voice similarity by drawing on acoustic features that have been found to be perceptually relevant. While Kelly et al. were able to show some correlation for their male speaker set, the study has several shortcomings with respect to the use of its results for the selection of foil voices in the voice parade context, as is the primary focus of the present study (see Section 1.3). In particular, recording conditions were unconstrained, accent, social background and age were not controlled, and listeners were not controlled for their native language. The present study improves on this work by conducting an experiment in which all of these factors are controlled.

There is one further study which warrants mention in the context of the present work, that of Park et al. (2018) which provides a comparison of human judgements and automatic assessments of pairs of voice samples for an ASR system incorporating phonetic features. However, this study is of speaker discrimination not speaker similarity, i.e. listeners are asked to judge same/different for each pair and results are compared with the automatic system's accuracy in same/different assessments, so the results are not directly comparable. In this study human and machine assessments differed, but there was a weak relationship between human- (MDS based on confidence judgements) and machine-determined spaces. The study finds different sources of speaker information being drawn on its human and machine results and notes the importance of acoustic phonetic features in accurate speaker discrimination judgements. Research is needed to establish relationships between human and automatic assessments of voice similarity and the extent to which inclusion of phonetic features in an ASR system can enhance such relationships. If the human perception of similarity is correlated with automatically identified similar voices, this will have an impact in several practical areas. It would allow film dubbing companies or games developers to choose local voice actors for their different characters that are closest to the voice of the original actor. Similarly, patients who have ‘lost’ their voice due to a degenerative disease or operation could choose a voice that is most similar to their own voice for synthesis. If recordings are available, it would be possible to use this technique to help determine the extent of similarity between their own voice and a synthesised voice. Further, correlation between human- and machine-assessed similarity of voices will have crucial implications for voice parade construction as discussed next in Section 1.3.

1.3 Voice parades

A potential application of automatically-determined assessments of speaker similarity lies in the construction of voice parades. A voice parade is an aural equivalent of a visual identity parade, i.e. a line-up of voice recordings which can be used to collect earwitness evidence for cases where a witness has heard a voice during the commission of a crime and typically not seen the perpetrator (due to dark surroundings, masking, blindfolding, etc.). A voice parade may be conducted when a suspect has been identified but there is no speech evidence, i.e. a voice recording, of the perpetrator's voice available.

In the UK, voice parades are prepared in accordance with the UK Home Office guidelines (Home Office, 2003, see also Nolan, 2003). An outline of the procedure is given in de Jong-Lendle et al. (2015). Regarding foil choice, the guidelines state that eight appropriate foil voices should be selected from a pool of at least 20 speakers "of similar age and ethnic, regional and social background as the suspect" (Home Office 2003: point 9), ensuring that "the accent, inflection, pitch, tone and speed of the speech used, provides a fair example for comparison against the suspect" (Home Office 2003: point 15).

To make sure that the selection of foil voices is fair, a paired comparison test has been introduced, as described by McDougall (2013a). This approach enables the phonetician to determine which eight speakers amongst a larger group of candidate foils are perceptually closest to the suspect speaker. The test involves asking a group of lay listeners to judge the similarity of every pairing of voices within the set containing the suspect and all candidate foils (cf. Rietveld and Broeders, 1991). Listeners rate the (dis)similarity of each pair of speakers on a 9-point Likert scale. These ratings are subjected to MDS. The eight candidate foils whose voices were judged most similar to the suspect's voice are then chosen as the foil-set for the voice parade.

Voice parades must be prepared bespoke to each individual case and setting up a voice parade involves extensive manual effort which is time-consuming and costly. Automating the selection of foils for a voice parade could facilitate and speed up the process, enabling voice parades to be carried out more straightforwardly and more often. Further, if it can be established that automatically selected foils are compatible with listener-based performance, an automatic system for selecting foils will provide an objective method for this task. A step towards this goal is to compare voice similarity ratings by human listeners with similarity scores produced by an automatic speaker recognition approach, in order to assess how far automatic approaches can be drawn on in replicating the perceived similarity of speakers.

1.4 Influence of language familiarity

An increasing number of cases involve the language spoken by the perpetrator being foreign to the earwitness, raising questions about the effect of language familiarity on the reliability of speaker recognition, and in particular whether speaker recognition performance differs between native-speaker listeners and listeners from another language background (Köster and Schiller, 1997: 18). Further, if the languages of speaker and listener are mismatched, the degree to which the languages are related might also have an impact on the speaker recognition performance (Köster and Schiller, 1997: 19). An overview of research investigating the language familiarity effect in voice parades and on the discrimination of speakers is given in Perrachione (2019: 6–8).

Research suggests that it can be expected that listeners with no or little knowledge of the target language will perceive the voices in a voice parade to sound more similar to each other while greater knowledge of the language allows differentiation on more perceptual levels (cf. Sherrin, 2015: 850). A study on the influence of language familiarity on voice similarity ratings was conducted by Fleming et al. (2014). Native speakers of Mandarin or English were asked to rate the speaker dissimilarity of pairs of English and Mandarin speakers using samples of duration 1250 ms. Although the phrases used for

the analysis were time-reversed to avoid a possible influence of language comprehension, both listener groups gave higher dissimilarity ratings to speaker pairs speaking the respective listener's native language and did not differentiate as much between the speakers of their foreign language.

San Segundo et al. (2016) explored multiple pairwise similarity ratings by 20 Spanish and 20 English native-speakers on short samples from Spanish twin pairs. They controlled the conditions for speaker similarity regarding dialect, age and F0, and listener groups were assumed to base their judgements on “holistic voice quality perception” (2016: 309) without an effect of language familiarity due to the short duration (~3 s) of the samples. Independent of listener language, twin pairs were perceived as more similar to each other compared with non-twins.

An analysis by Perrachione et al. (2019) replicated Fleming et al. (2014) and evaluated a possible language familiarity effect on similarity ratings for read sentences in Mandarin and English, played to listeners in both forward and reversed conditions. Listeners were native speakers of Mandarin or English. While high correlations between the voice similarity ratings of the listener groups and forward and reversed speech were found, the observed effect of language familiarity was minimal, in contrast with Fleming et al. (2014) results.

The language familiarity effect appears to play a role not to be disregarded in speaker recognition performance of human listeners. The few studies available of its effect on voice similarity ratings on the other hand suggest that it might not be as important here as for human speaker recognition. Further research is needed, however, including investigation of the assessment of voice similarity by native speakers of different languages as compared with an automatic system, as is undertaken in the current study.

2 Research questions

The present study aims to investigate the relationship between listener-perceived voice similarity and speaker similarity determined by an automatic system. The study compares demographically controlled voices, that is, speakers of the same sex, age group, and accent background.

Additionally, the influence of the first language of listeners on the judgement of voice similarity is explored. An online listener test is employed, with most participants being recruited from England and Germany. Since most listeners describe themselves as native speakers of either (British) English or German, the presence of language-specific patterns amongst the ratings of these two listener groups will be investigated. Assuming that German listeners will report an overall good knowledge of the English language, and given the high degree of relatedness of the two languages, it is anticipated that the perceived similarity of voice pairs will be comparable between the two groups.

3 Method

3.1 Voice database

Stimuli were created using the DyViS ('Dynamic Variability in Speech') database which consists of the voices of 100 male speakers aged 18–25 years with a Standard Southern British English (SSBE) accent (Nolan et al., 2009; Nolan, 2011). This database of accent- and demographically-matched voices was chosen in order to replicate as closely as possible the real case situation of a voice parade in which a phonetician selects foil voices from amongst a collection of interview recordings of speakers with the same demographic background as the suspect, as noted in Section 1.3. The male/18–25 years/SSBE accent demographic combination was selected here as a proxy for a homogeneous set of speakers, but any fixed combination of demographic features and recording conditions could have been chosen.

For the automatic comparison, studio quality recordings from Task 2 (spontaneous speech) and Task 3 (read speech) from the DyViS database were used. Task 2 recordings were also used to create stimuli for the listener experiment.

3.2 Automatic experiment

The automatic experiment was conducted using the VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence) speaker recognition system (Alexander et al., 2016; Kelly et al., 2019). VOCALISE supports x-vector and i-vector speaker recognition frameworks, along with classic GMM-based approaches. All choices of framework can operate with either spectral Mel frequency cepstral coefficients or 'auto-phonetic' features. These auto-phonetic features are considered to be more perceptually relevant than frequency envelope features like Mel-frequency cepstral coefficients. VOCALISE offers the capability of being able to extract and use these features with different algorithms typically used in automatic speaker recognition like GMMs, i-vectors and x-vectors. Measurable phonetic features such as formants and pitch are typically used by phoneticians in forensic phonetic casework to compare speakers. To the best of the authors' knowledge VOCALISE is one of the only commercial forensic automatic speaker recognition systems that provides the capability of automatically extracting these features, and modelling them using ASR algorithms.

For this paper, VOCALISE was used in i-vector PLDA (Probabilistic Linear Discriminant Analysis) mode, with auto-phonetic features (Alexander et al., 2016). To compare speakers, VOCALISE first extracts auto-phonetic features from each speech sample. The auto-phonetic mode was considered a good choice because phonetic features have been shown to be linked with perceived voice similarity (e.g. Nolan et al., 2011, McDougall, 2013). In this experiment, the adapted auto-phonetic mode from Kelly et al., which extracts "F0, F1-F4, semitones of F0, along with first derivatives" (2016: 1567) is used. Speaker modelling is based on a Universal Background Model (UBM) with 256 components, and a 200-dimensional Total Variability (TV) matrix (Kelly et al., 2016: 1568). i-vectors are subjected to Linear Discriminant Analysis (LDA) to reduce the number of dimensions to 100 and ensure greater speaker separability. Finally, 100-dimensional

PLDA is used to model intra-speaker and inter-speaker variability and calculate the likelihood of two i-vectors coming from the same versus different speaker(s). For each pair of i-vectors that is compared, a score is output (Alexander et al., 2016).

The stimuli for the automatic comparison had a duration of 2–6 min. All Task 2 and Task 3 studio recordings from 100 speakers were compared against each other. This produced a matrix of 9900 different-speaker and 100 same-speaker comparison scores. Cross-validation score calibration (via linear logistic regression) was applied to the scores using Bio-Metrics[2] in order to normalise their numerical range (Pigeon et al., 2000).

3.3 Stimuli for the listener experiment

Due to the constraints of the listener experiment and the trade-off between more robust results and listener fatigue, it was decided to use VOCALISE comparison scores calculated for the full set of 100 DyViS speakers to inform the selection of a smaller more manageable subset of ten speakers for the listener test. This subset of speakers was selected based on the distribution of their VOCALISE comparison scores across four quartile 'bins'. This approach allowed speakers with an even spread of low to high comparison scores to be selected, and outlier speakers that yielded overall very high or very low comparison scores to be excluded.

The experimental design involved selecting a set of speakers that yielded high VOCALISE comparison scores for same-speaker comparisons and a range of VOCALISE comparison scores from low to high for different-speaker comparisons. It was assumed that speakers who yielded an even spread of different-speaker comparison scores (across four quartile 'bins') when compared against the full 100 DyViS speaker set would also exhibit a relatively even spread of different-speaker comparison scores within the final subset of ten speakers.

The full set of different-speaker VOCALISE comparison scores was divided into quartiles. Excluding the same-speaker comparison, each individual speaker's VOCALISE comparison scores were ranked from 1 (lowest score) to 99 (highest score). For each speaker, the distribution of these ranks across the quartiles was examined and the standard deviation of the number of ranks within a quartile was calculated. These standard deviations formed a measure of how balanced each speaker's VOCALISE comparison scores were, i.e. a speaker with a high standard deviation had been calculated to yield an above average number of VOCALISE comparison scores from within one certain 'bin' and fewer in the other 'bins'. Such a speaker could, for example, yield very high or very low overall VOCALISE comparison scores in most comparisons. Standard deviations ranged from 1.48, with spk048 and spk059 having 23 to 27 comparison scores in each of the four 'bins', to 30.22 for spk074 with 77 of 99 different-speaker VOCALISE comparison scores in the 'bin' containing the lowest scores. The VOCALISE comparison scores of the ten speakers with the lowest standard deviations were extracted and allocated to the respective 'bin' in order to assess whether scores from all 'bins' were present in roughly similar numbers. One speaker yielded overall very high VOCALISE comparison scores

[2] Bio-Metrics 1.8 performance metrics software, Oxford Wave Research Ltd., https://www.oxfordwaveresearch.com/products/bio-metrics, accessed 30 January 2020.

(i.e. most of their scores were in the fourth 'bin') and was swapped with the speaker with the next lowest standard deviation. The final subset of ten speakers will be referred to as spk017, spk027, spk029, spk036, spk054, spk059, spk060, spk079, spk093, and spk095, where the numbers refer to the respective DyViS participant numbers.

A pilot study with eight listeners showed that it was hard for listeners to compare voices only when also confronted with the mismatch in speaking style that comes with comparing spontaneous speech from Task 2 against read speech from Task 3. Based on the assumption that spontaneous speech would be encountered more often in a voice parade scenario than read speech, it was decided that only Task 2 files would be used in the final listener experiment.

From each selected speaker's Task 2 recording, samples of continuous speech of 3 to 5 s duration were manually extracted. In order to use samples that reflect the scores of the longer file comparisons from Task 2 and 3, all short samples from Task 2 were compared against each other in VOCALISE. The quartiles based on these scores were calculated using the same method as was used above for the longer samples to ensure that the 3–5 second samples yielded the same quartile results as the longer samples. Eleven samples per speaker were selected so that the resulting listener test stimuli always contained new samples of the speakers concerned. 45 different-speaker and ten same-speaker comparisons were constructed. The selection of the samples was based on whether the respective comparison scores belonged to the same quartile as the score resulting from the longer recording comparisons.

The original comparison scores of the ten selected speakers (pre cross-validation calibration) were then re-calibrated using the 90 deselected speakers. Calibrating the scores of the selected speakers with an independent set of speakers in this way gives a fairer representation of a scenario in which one would apply pre-determined calibration parameters to new sets of potential foil speakers. To apply this calibration, Task 2 and Task 3 recordings from the 90 deselected speakers were compared against each other using VOCALISE as previously described. This was repeated for the ten selected speakers. Calibration parameters derived from the scores of the 90 speakers were then applied to calibrate the score matrix of the set of ten speakers using Bio-Metrics[3]. The VOCALISE scores resulting from the Task 2 sample 1 versus Task 3 sample 2 comparison and the Task 2 sample 2 versus Task 3 sample 1 comparison were averaged for each speaker pair.

In a nutshell, ten of the DyViS speakers were selected to form a subset for the listener experiment based on the distribution of their VOCALISE comparison scores across four quartile 'bins' with the aim to get close to an even spread of high to low VOCALISE comparison scores. Outliers that had yielded overall very high or very low comparison scores – meaning they could be either very easily confused or distinguished using the auto-phonetic mode in VOCALISE – were excluded. Only spontaneous speech from Task 2 was used for the listener experiment and eleven stimuli per speaker with a duration of 3 to 5 s each were selected to make up 45 different-speaker and ten same-speaker comparisons. Recalibration was applied to the original comparison scores of the

[3] Bio-Metrics 1.8 performance metrics software, Oxford Wave Research Ltd., https://www.oxfordwaveresearch.com/products/bio-metrics, accessed 30 January 2020.

selected speakers to approximate a realistic scenario in which a new set of foils would be calibrated using pre-determined parameters.

3.4 Listener experiment

To avoid listener fatigue and participants abandoning the assessment early, the experiment was designed to be as short as possible and only to collect ratings for comparisons which were absolutely necessary. In order to assess the perceptual distances between all speakers, every speaker had to be compared at least once with every other speaker, assuming that e.g. spk017 vs. spk027 would yield the same rating as spk027 vs. spk017. Hence, instead of as many as 90 different-speaker comparisons, 45 were considered sufficient in the perceptual distance experiment. Same-speaker comparisons were included in order to confirm whether listeners indeed perceived samples from the same speaker as most similar to each other. The order of samples within a speaker comparison and the position of a comparison within the experiment were randomised but remained the same for each listener.

An online experimental interface was used for the perceptual task in order to maximise participant numbers. The assessment was divided into a training phase and a testing phase. The training phase served to familiarise the participants with the format of the experiment and the range of voices, and to allow them to practise making judgements. It consisted of five comparisons made up of two short clips each. The samples were drawn from the ten selected speakers and were comparable to those used in the testing phase. The participants were asked to rate the overall voice similarity of each given pair in a snap reaction on a scale from 1 (very dissimilar) to 9 (very similar), ideally while ignoring accent, speaking style, and speech content. The participants were also told that they would encounter male voices only. The speed at which listeners could proceed lay in their own hands. To submit each response, listeners were required to click a ⟨Submit⟩ button at the end of the comparison row. After clicking that button, the row was disabled and listeners could not go back to the respective comparison to re-listen or to change their rating. The testing phase followed the same schema as the training phase. Fig. 1 exemplifies the experimental interface by showing the first two comparisons.[4]

Figure 1. Comparisons in the online assessment (desktop version using Firefox) with disabled row No. 1 showing a submitted score.

[4] The online assessment was made accessible via the Oxford Wave Research website: https://www.oxfordwaveresearch.com/voicesim2019/, last accessed 30 January 2020.

3.5 Listeners

A total of 112 listeners participated in the online experiment over the course of five weeks. The results from six individuals were excluded due to technical problems, a hearing impairment, or an unusual distribution of the participant's ratings (e.g. the participant's later responses only using one value on the scale, indicating that the participant was no longer engaging properly with the task). One of these listeners rated 25 consecutive comparisons (comparisons No. 31–55) as '5', indicating that they lost interest in the experiment. Two other listeners were very extreme outliers regarding the frequency with which they chose '1 – very dissimilar' on the rating scale. These listeners rated 40 and 35 out of 55 comparisons as '1' and, therefore, appeared to have rated the voice pairs rather regarding same versus different instead of similar versus dissimilar, so their results were excluded.

106 listeners – 45 male, 58 female and 3 other – were thus included in the analysis. The participants were aged between 18 and 63 years with an average age of 26.46 years (SD = 8.79 years) and a median of 24 years. In total, 16 different languages were indicated as participants' first languages. German and English were the most frequent first languages, with 46 and 41 listeners respectively. The remaining listeners' first languages included Chinese (n = 4), French and Greek (n = 2 each) and Nepali, Frisian, Indonesian, Hindi, Lithuanian, Finnish, Hungarian, Thai, Swedish, Yoruba, and Turkish (n = 1 each).

According to their self-reported accent or dialect, 39 listeners were speakers of British English (BE). These are relevant for the exploration of rating differences between German listeners and listeners who are substantially familiar with BE which was spoken in the comparison samples. The accent of a huge proportion of English listeners could be summarised as Standard Southern British English (SSBE); these included the descriptions "RP" (Received Pronunciation), "Queen's", "BBC English", "SSBE" and "none". Other English accents and regions mentioned were, e.g. "Yorkshire", "Lancashire", "North Wales", "Coventry", "London" or "Home Counties". The majority of German listeners (n = 34) reported speaking standard German, seven indicated Bavarian, three Hesse and one each Alemannic and Saxonian as their respective dialects.

Participants were asked to report their proficiency in English. In total, 43 participants indicated they were native or bilingual speakers of English. This includes speakers of, e.g. American and Australian English. The proficiency of non-native speakers of English was either "Basic", "Good", "Very good" or "Excellent". The majority of non-native English speakers reported they had a very good proficiency in English (n = 32) or excellent proficiency (n = 18). Twelve listeners stated that they had good proficiency, and one listener estimated their proficiency to be basic.

Roughly two thirds of the listeners completed the voice similarity assessment on their computer or laptop, another third used their smartphone or tablet. About 75% of the listeners used in-ear or over-ear headphones and roughly 21% used their internal speakers. The remaining listeners used loudspeakers. None of the included listeners reported a hearing impairment and all of them reported being in a quiet environment when completing the experiment.

3.6 Evaluation and analysis of the data

In making their voice similarity assessments, listeners did not use the rating scale consistently with each other, and some did not use its full range. For each speaker pair in the listener test, any listener ratings in the bottom and top 5% of the spread of responses to that comparison were removed, then a mean of the remaining 90% of scores was taken, i.e. a 5% trimmed mean, to get a first impression of the correlation between calibrated VOCALISE scores and listener ratings.

A more detailed examination of the untrimmed data was then carried out using Weighted MDS, a type of MDS which, as well as calculating the usual stimulus space, provides a participant space that indicates the weighting given by each participant to each dimension in the common stimulus space, taking into account individual differences in the listeners' use of the Likert scale (see Giguère, 2006). This was executed using the statistical software SPSS. Since Weighted MDS is based on multiple comparison matrices, the technique was applied to the matrices of all included listeners (n = 106), as well as to those of BE participants and German participants separately. There were fewer BE participants (n = 39) than German participants (n = 47), so a subset of 39 German participants was chosen randomly to provide a balanced dataset for the BE-German analyses. The similarity scale in the voice similarity assessment was used such that higher numbers indicated higher similarity. To aid interpretation, the ratings were inverted for each listener before subjecting the data to MDS (cf. Giguère, 2006: 29). In line with recommendations from Kruskal (1964: 1), all ratings were subtracted from the number ten to invert them, i.e. to convert the similarity ratings to dissimilarity ratings (or distances).

Since the VOCALISE scores consisted of a single matrix, Weighted MDS was not appropriate. For further processing, the uncalibrated scores first were converted into distances to be comparable to the listener MDS. The scores were scaled from 1 to 9 and then inverted to represent dissimilarity now rather than similarity, as is the approach in MDS. Following this, Principal Component Analysis (PCA), a basic type of MDS (Abdi, 2007: 1), was applied.

4 Results

4.1 Relationship between human listener and machine ratings

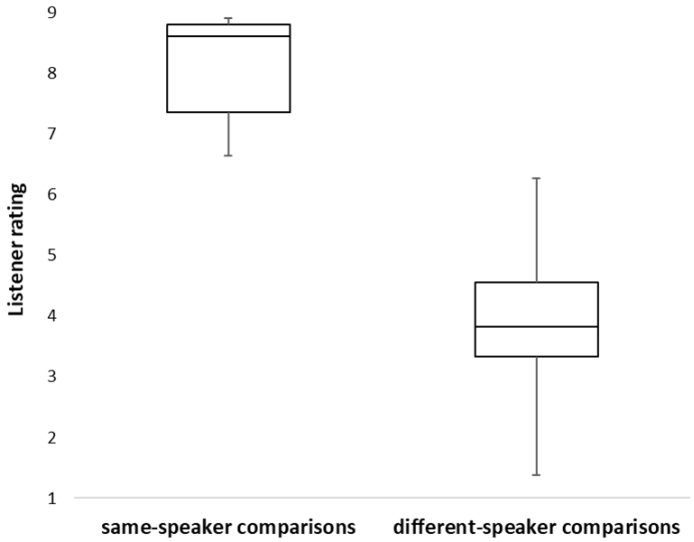

Figure 2. Box plot displaying the 5% trimmed mean listener ratings for same- and different-speaker comparisons (Listener rating 1 = very dissimilar, 9 = very similar).

To demonstrate the distribution of listeners' responses to same-speaker and different-speaker comparisons on the Likert scale, boxplots showing the 5% trimmed mean listener ratings are given in Fig. 2. The box plots in Fig. 2 reveal that the median of the listener ratings for same-speaker comparisons (MED = 8.60) is substantially higher than that of different-speaker comparisons (MED = 3.82). For same-speaker comparisons, the 5% trimmed mean of ratings ranges between 6.64 and 8.91, whereas for different-speaker comparisons the minimum is at 1.38 and the maximum is at 6.26. For same-speaker comparisons, there are a range of responses: they are not judged consistently with a perfect 9, showing that listeners are treating same-speaker pairings in a similar conceptual/perceptual way to different-speaker pairings.

Fig. 3 displays a scatter plot with the 5% trimmed mean of the listener ratings of voice similarity on the horizontal axis and the calibrated VOCALISE scores on the vertical axis. The 45 different-speaker comparisons are shown as squares, while the ten same-speaker comparisons are triangles. As can be seen in the plot, different- and same-speaker comparisons do not overlap and are clearly separated by both the automatic approach and by the listeners. Different-speaker comparisons produced lower voice similarity scores than same-speaker comparisons for both listener judgements and the automatic approach. The comparison of spk060 to spk017 yielded the lowest similarity scores, both for listeners (5% trimmed mean = 1.38) and for the machine approach (calibrated VOCALISE score = −8.63). All same-speaker comparisons received high scores on both axes. Three speakers (spk027, spk036, spk054) have received comparatively very high voice similarity scores for their within-speaker comparisons by the automatic approach, but their voices were not judged by the listeners as being as similar to themselves as the rest of the speakers' voices.

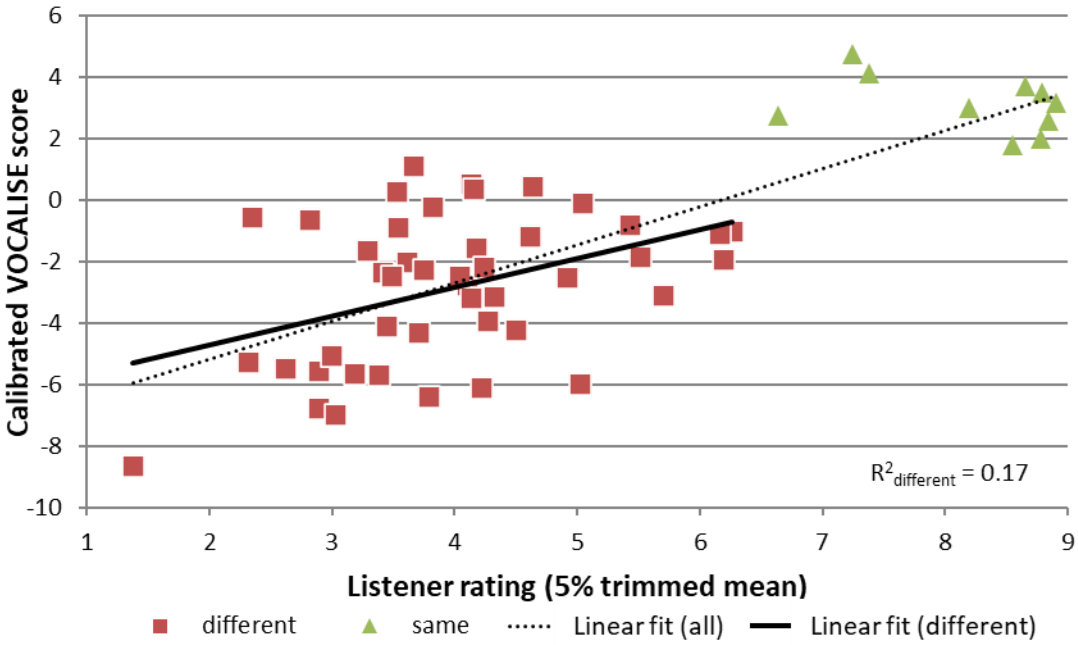

Figure 3. Scatter plot displaying the relationship between calibrated VOCALISE scores and the 5% trimmed mean of the listener ratings (Listener rating 1 = very dissimilar, 9 = very similar).

Linear regression was applied to the different-speaker comparison scores using Excel, and the resulting regression line ($R^2_{different}$ = 0.17, $R_{different}$ = 0.42) is shown in the plot as a solid line. For comparison, the linear regression line for all comparisons (both same-speaker and different-speaker) is shown in the plot as a dotted line. Correlating scores from different-speaker comparisons using Spearman rank correlation results in ρ = 0.34* (p = 0.02, 2-tailed). This result confirms that there is a statistically significant positive correlation between listener ratings and VOCALISE scores for different-speaker comparisons that is broadly linear.

Listeners used the scale from 1 to 9 in different ways and not always the full range. Most listeners used the extreme ends of the scale at least once. Some, however, focussed rather on the extremes and used few intermediate ratings to judge the similarity of voice pairs in a comparison. Other listeners did not make use of the full scale but rather tended to use the upper or the lower half.

A method that takes listeners' individual approaches to the use of the rating scale into account is Weighted MDS (also INDSCAL – Individual differences scaling) (cf. Giguère, 2006: 32). Weighted MDS uses distance matrices of multiple pairwise comparisons as input, in this case those supplied by 106 listeners through their ratings. An analysis based on Weighted MDS aims to find a lower dimensional representation of the distances between speakers as perceived by the listeners (cf. Cox and Cox, 2000: 1). Only different-speaker comparisons are considered in the MDS analysis.

By inverting the original ratings in the listener matrices, high ratings denote low similarity, whereas low ratings correspond to high similarity. This can be observed as large Euclidean distances between data points of speakers perceived as dissimilar, whereas data points of speakers perceived as similar appear closer to each other in an MDS analysis.

In this Weighted MDS analysis, an analysis with five pseudo-perceptual dimensions was chosen, as it provided a "fair" fit according to the stress thresholds proposed for goodness of fit (cf. Kruskal, 1964: 3) for the given listener similarity ratings (stress = 0.14, R^2 = 0.20). The first two MDS dimensions which account for the greatest amount of variance successively can be represented on a scatter plot to enable visualisation to some extent of the (dis)similarity relationships between the ten speakers. Such a scatter plot is shown in Fig. 4 with MDS dimension 1 on the horizontal and MDS dimension 2 on the vertical. For comparison, Fig. 5 gives a scatter plot of the first two of the five PCA dimensions produced by the VOCALISE scores.

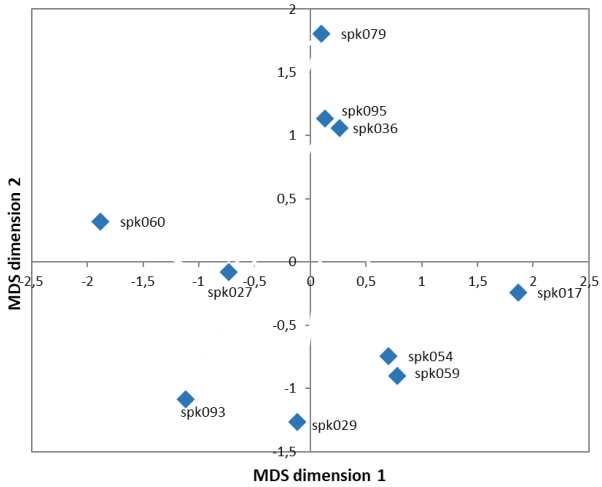

Figure 4. Plot of the first two dimensions produced by MDS based on the ratings from 106 listeners.

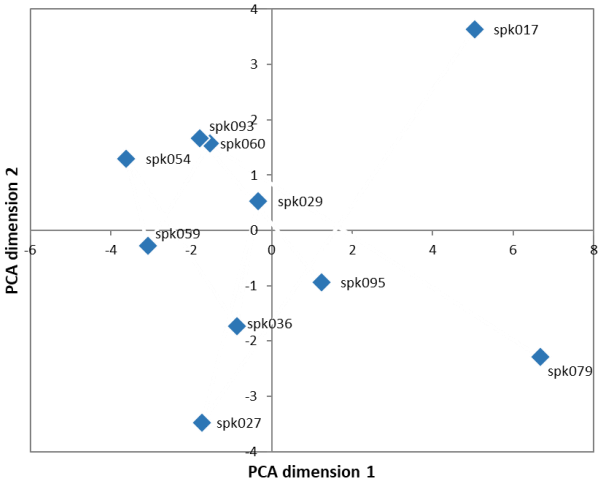

Figure 5. Plot of the first two dimensions produced by PCA based on VOCALISE scores.

Several similarities between the two plots can be observed. Spk095 and spk036 are the closest pair in the MDS plot. These two speakers are also close neighbours in PCA dimension 2, and do not vary greatly on PCA dimension 1. Similarly, the cluster of spk054 and spk059 in the MDS plot can be found near each other in the PCA plot, where the speakers are very similar in dimension 1 and vary slightly in dimension 2. Spk017 and spk079 seem to be quite dissimilar from any of the other speakers based on the VOCALISE scores. The same tendency, though not as strong as in the PCA plot, can be found in the MDS plot. Also, spk079 being closer to spk095 and spk036 in the MDS plot is reflected in the PCA plot, where spk079 is similar to that speaker pair in the second dimension. The other speakers do not seem to build any striking clusters in the MDS plot; however, they are all rather similar in the first dimension and vary in their degree of similarity in the second dimension. In PCA, the other speakers are relatively slightly more similar in dimension 1 than in MDS, and differ comparatively more on dimension 2 but overall loose similarities between the plots can also be found for those speakers.

The results in the scatter plot in Fig. 3 revealed a positive and significant correlation in aggregate. In order to get a clearer picture of the speaker-specific (dis)similarities within MDS and PCA, the Euclidean distances between all speakers were calculated using the five dimensions for each of the MDS and PCA results using the following formula: $d(spk_p, spk_q) = \sqrt{\sum_{i=1}^{5}\left(dim_{i_{spk_q}} - dim_{i_{spk_p}}\right)^2}$. Spearman rank correlation was used to further explore the relationships between the MDS and PCA analyses with respect to the

speakers' Euclidean distances to other speakers in the two multidimensional spaces. The results are listed in Table 1. A significant correlation between the Euclidean distances in MDS and PCA was found for spk029 (ρ = 0.76, p = 0.01). For all other speakers, the equivalent correlation is not significant. Nevertheless, relatively high correlations are shown by spk054, spk059, spk093 and spk095 whose correlation coefficients are greater than ρ = 0.3.

Speaker	Spearman's rho (2-tailed)
spk017	-0.22
spk027	0.30
spk029	0.76* (*p = 0.01*)
spk036	-0.01
spk054	0.31
spk059	0.37
spk060	0.13
spk079	0.13
spk093	0.47
spk095	0.41

Table 1. Spearman rank correlations of each speaker's Euclidean distances in MDS and PCA (*significant correlation at the 0.05 level, two-tailed).

4.2 Differences in similarity ratings of native German and British English listeners

The first language of listeners may have an influence on the voice similarity ratings of the native speakers of BE in this experiment. In this case, most listeners' native language was either German or (British) English, therefore, MDS was applied to the subsets of matrices from German listeners and BE listeners. In both cases, the five-dimensional analysis was selected (BE: stress = 0.14, R^2 = 0.24; German: stress = 0.14, R^2 = 0.22). Fig. 6 presents a scatter plot of the first two dimensions from each MDS analysis, the one for BE listeners on the left and for German listeners on the right.

The plots show some differences in the distances between the clustering of speakers across the analyses, however, some loose groups can be identified. Circled in light-grey dashes in the bottom left quadrant of the BE MDS plot are spk059, spk029, and spk054. Although German listeners seem to judge the voices of spk029 and spk059 as more similar to each other than each of them compared to spk054, they are all still relatively close together in the bottom left quadrant of the BE MDS plot. A group of voices that German listeners rate as very similar, those of spk060, spk027, and spk093 (black solid ellipse), is at least judged very similar in dimension 2 by BE listeners. Spk017 (light-grey solid ellipse) seems to be rated as somewhat dissimilar from other speakers by BE listeners, whereas German listeners perceived this speaker as quite similar to spk079 and spk036 (dark-grey dashed ellipse). However, this group and spk017 can still be found in the same quadrant in the BE listener MDS plot. Since the orientation of MDS is undefined, and the MDS plot can thus be legitimately rotated, here if one were to rotate the German listener MDS plot by 90° to the left, one would find a distribution of the speakers quite

similar to the BE listener MDS plot. As the languages are closely related and most German listeners reported a good to excellent proficiency in English, this is not a surprising result. Overall, the MDS plot based on the ratings from all 106 listeners mostly reflects similar relationships to those found in the separate plots for German and BE listeners. Differences could be due to combining the ratings from both German and BE listeners, together with ratings from the listeners with another language background. An underlying reason might be that foreign-language speakers of English lack the ability to perceive certain sociophonetic or dialectal features to which native speakers are sensitive.

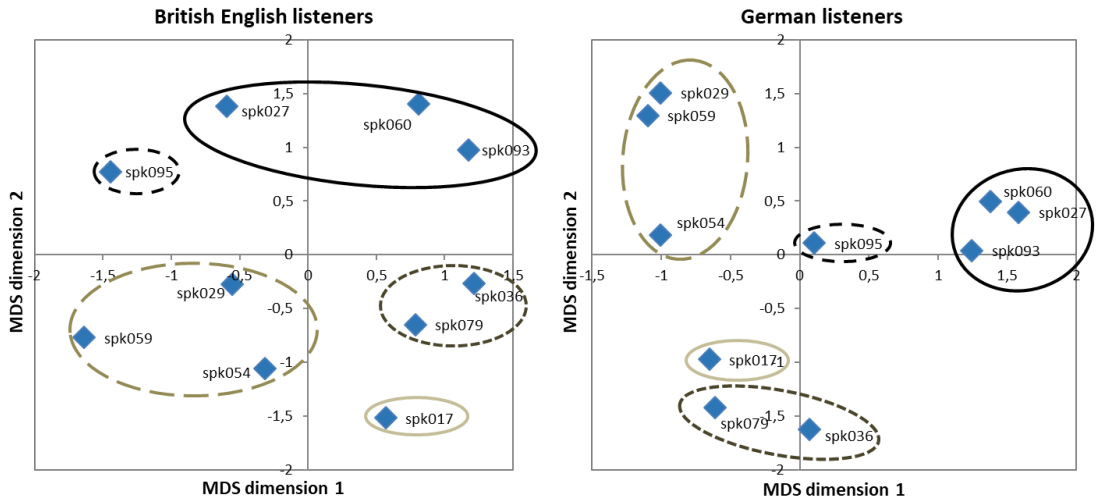

Figure 6. MDS plots in two dimensions based on matrices from British English listeners (left) and German listeners (right). Ellipses are explained in the text.

Comparing the separate MDS plots (Fig. 6) with the PCA plot (Fig. 5), many similarities can be observed. Spk054 and spk059 are perceived as rather similar to each other according to both the MDS plots and they are close together in the PCA plot. Similar relationships also exist for spk093 and spk060. While spk017 and spk079 can be found relatively close together in the same quadrant within the MDS plots, they are similar on PCA dimension 1 as well. Also relatively similar in all plots are spk029 and spk59 and close on at least one dimension within any of the plots are spk027 and spk093.

To compare relationships between the ratings of the BE and German listener groups further, the Euclidean distances of each speaker to all other speakers were calculated individually using the five dimensions produced in each of the MDS analyses of ratings by BE listeners and German listeners. The correlations between the distances for each speaker to all other speakers in the two MDS analyses were explored using Spearman rank correlation as shown in Table 2. All speakers yielded positive correlations, and six of these were significant or highly significant.

Speaker	Spearman's rho (two-tailed)
spk017	0.77** (*p = 0.01*)
spk027	0.92** (*p = 0.00*)
spk029	0.33
spk036	0.61
spk054	0.22
spk059	0.16
spk060	0.73* (*p = 0.02*)
spk079	0.72* (*p = 0.02*)
spk093	0.81** (*p = 0.01*)
spk095	0.89** (*p = 0.00*)

Table 2. Spearman rank correlations of each speaker's Euclidean distances in MDS based on BE and German listeners respectively. (*significant correlation at the 0.05 level, **significant correlation at the 0.01 level, two-tailed).

The procedure was repeated separately for the distances based on the MDS analyses of BE and German listener ratings and for the distances based on the PCA dimensions. The results of the Spearman rank correlations are shown in Table 3. All but one of the correlation coefficients between the PCA and the German MDS analysis, are positive, with spk029 showing the strongest positive correlation coefficient (ρ = 0.588). However, no significant correlations are observed here. By comparison, in the PCA and BE MDS analysis there are several larger correlation coefficients present, including two significant ones (spk054, ρ = 0.673*, p = 0.033; spk095, ρ = 0.709*, p = 0.022). Yet there are three negative correlation coefficients (spk017, spk036, spk060). Overall, the BE listeners seem to agree better with the automatic distances than the German listeners.

British English listeners		**German listeners**	
Speaker	**Spearman's rho (two-tailed)**	**Speaker**	**Spearman's rho (two-tailed)**
spk017	-0.07	**spk017**	-0.29
spk027	0.25	**spk027**	0.31
spk029	0.06	**spk029**	0.59
spk036	-0.19	**spk036**	0.10
spk054	0.67* (*p = 0.03*)	**spk054**	0.02
spk059	0.26	**spk059**	0.44
spk060	-0.06	**spk060**	0.18
spk079	0.43	**spk079**	0.15
spk093	0.42	**spk093**	0.33
spk095	0.71* (*p = 0.02*)	**spk095**	0.49

Table 3. Spearman rank correlations of each speaker's Euclidean distances in PCA and MDS based on BE listeners (left) and German listeners (right) respectively (*significant correlation at the 0.05 level, two-tailed).

5 Discussion

Displaying listener and machine ratings in a scatter plot showed a pattern of higher ratings for same-speaker and lower ratings for different-speaker voice comparisons for both

human listeners and the automatic approach. Focussing on different-speaker pairs, a statistically significant, broadly linear relationship was found between listener and machine ratings. This indicative finding supports further exploration of the use of automatic systems for speaker similarity assessment. Given sufficient agreement between scores from an automatic approach and listener judgements, the automatic approach would provide an objective way to select foil voices for voice parades. Score calibration would allow for direct interpretation of the numerical values as similarity measurements, e.g. to compare similarity directly across different sets of foils. Additionally, a similarity score range that is acceptable for voice parades could potentially be defined. Ideally, a forensic phonetician would be able to use a suspect's speech sample in an automatic voice similarity comparison to preselect a set of similar voices from a pool of voices followed by a significantly shorter auditory analysis than would be necessary without this preselection. Given the availability of appropriate, homogeneous speaker databases, time spent on an auditory analysis could be reduced further.

Visual comparison of the MDS and PCA plots displaying the speaker similarity relationships as perceived by listeners and determined by a machine showed many parallels. In order to compare these parallels numerically, the Euclidean distances between all pairs of speakers were calculated using five dimensions of MDS and PCA. Spearman rank correlation analysis of each speaker's Euclidean distances to all other speakers within MDS and PCA revealed positive correlations for a number of speakers. While many parallels were observed, the ratings of the listeners and the automatic approach were not always well aligned. This raises questions about what listeners might further take into account that is not (yet) included in the analysis by the automatic approach. For example, one listener commented that a speaker in the experiment had a lisp. A preliminary auditory analysis revealed spk017 as that particular speaker. This might contribute to an explanation as to why this speaker was isolated from the group in the MDS and PCA analyses. As McDougall (2013b) has previously shown, various voice quality features (including a fronted or advanced tongue position) can play a role in listener-based voice similarity judgements. A future analysis of voice quality and acoustic correlates may shed further light on the present results.

A comparison of BE and German listeners' voice similarity ratings separately was conducted to explore a potential effect of the first language of listeners on the ratings and to investigate which group's ratings were more similar to the ratings of the automatic approach. Spearman rank correlation of each speaker's distances to all other speakers in MDS for BE listeners and German listeners resulted mostly in positive correlations, some of which were significant or highly significant. Although some previous research has shown a language familiarity effect in voice similarity ratings (e.g. Fleming et al., 2014), such an effect did not seem to play a huge role in the present experiment whose findings were instead more consistent with those of Perrachione et al. (2019), perhaps due to the fact that the German listeners almost all indicated having a "good" or "very good" proficiency in English. Furthermore, the sample duration was very short with roughly four seconds per sample, and hence the overall impression of voice quality might be more important than the first language of the listeners. This would support findings by San Segundo et al. (2016) which showed comparable ratings by naïve English and Spanish listeners on short samples from a homogeneous group of Spanish speakers.

Comparing the Euclidean distances of the MDS analyses based on BE and German listener ratings to the PCA analysis based on the ratings from the automatic approach, it was the BE listener ratings which yielded the greater number of significant correlations with the automatic output. The analysis based on German listener ratings did not yield any significant correlations, but did produce some positive correlations.

One explanation for the results differing between the two listener groups might be that BE listeners perceive finer differences than German listeners and that these differences are also taken into account by the automatic approach. These differences could concern, for example, vowel qualities which have an influence on formant measurements. In a situation like the present where the speakers are heavily controlled for their gender, age, accent, and social background and the conversation content revolves around the same topic, the speakers and their speech are already quite similar overall. While holistic voice quality might be perceived similarly by BE and German listeners, the former might be more susceptible to slight variations of (socio)phonetic and linguistic cues between the speakers. It is possible that sociophonetic variation has an influence on formant measurements which are reflected in the ratings from the automatic approach. This, however, needs further investigation. Nevertheless, the individual variation in the correlation coefficients shows the need for further research with more speakers.

Another factor to consider is that the linguistic content of the samples was not identical, so listeners were unable to make direct linguistic comparisons. It is possible that listeners may have taken linguistic features (‘what was said’) into account when rating the speakers’ samples even when asked to rate the similarity of ‘voices’. This could have led to some discrepancies between human listener and machine ratings. Thus, additionally to voice quality, (socio)phonetic and linguistic information relating to the speakers may also have contributed to the positioning of the speakers along the pseudo-perceptual dimensions determined by the listener ratings and to discrepancies between listener and machine ratings.

6 Conclusion

In this study, the assessments made by human listeners of the extent of similarity amongst pairs of voice samples were correlated with the corresponding similarity scores obtained from an i-vector system using auto-phonetic features on a small scale. Relationships between the two types of similarity were demonstrated with correlation analysis and various analyses based on Weighted MDS and PCA. Separating responses from English and German native speaker groups showed similar patterns of a positive relationship between listener and automatic assessments, but with English listener assessments achieving higher levels of correlation with the automatic output than those of the German listener group. Future work will investigate this phenomenon for larger groups of speakers, and for different accents of English and other languages. Additionally, the comparison of female speakers’ voices and possible sex-specific effects must be addressed in future experiments when appropriate databases are available. The effectiveness of automatic systems in reflecting listeners’ voice similarity judgements requires additional testing.

Further work will also entail an evaluation of the acoustic-phonetic features which underlie the pseudo-perceptual dimensions generated by the MDS analysis and investigation of whether the specific words and phrases contained in speech samples play a role in how listeners make judgements of voice similarity. In order for automatically retrieved voice similarity ratings to align more with what listeners perceive as similar, improvements to the automatic approach based on phonetic features that are found to play a role must be further explored.

Many practical implications of this work for voice parade construction await further exploration. The findings of this study pave the way for future developments towards an automatic approach for selecting foil voices for voice parades. Calibration of automatically generated scores would enable similarity comparisons to be made across different sets of candidate foils, and a threshold for an acceptable extent of similarity between foils and a suspect voice could potentially be identified. Further, the ability to identify perceptually similar speakers using an automatic approach can contribute to improvements in voice casting and voice banking application development.

7 Acknowledgements

The authors are grateful to Gea de Jong-Lendle of Philipps-Universität Marburg for her initial involvement in establishing the project and her continued support throughout. Thanks to Samuel Kent from Oxford Wave Research for his work on the web-based interface that was used in the listener experiment and to Oscar Forth for his modifications of the VOCALISE software to allow for these comparisons. Thanks are also due to the experimental participants who gave freely of their time.

8 References

Abdi, H. (2007). Metric multidimensional scaling (MDS): analyzing distance matrices. *Encyclopedia of Measurement and Statistics*, 1-13.

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.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research PRPF*, *74*(1), 110-120.

Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. Chapman and hall/CRC.

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, *111*(38), 13795-13798.

Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology*, *2*(1), 27-38.

Home Office (2003). Advice on the use of voice identification parades. UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group, Police Leadership and Powers Unit. Available at: www.homeoffice.gov.uk/about-us/corporate-publications-strategy/home-office-circulars/circulars-2003/057-2003/ (29.01.2019).

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015, August). Voice lineups: a practical guide. In *18th International Congress of Phonetic Sciences. Glasgow, Scotland* (pp. 10-14).

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 Proceedings of INTERSPEECH 2016, San Francisco, 1567-1568.

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.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, *4*(1) 18-28.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27.

Lindh, J. (2009). Perception of voice similarity and the results of a voice line-up. In Proceedings of *FONETIK 2009, The XXII Swedish Phonetics Conference*, Department of Linguistics, Stockholm University. 186-189.

Lindh, J., & Eriksson, A. (2010). Voice similarity - a comparison between judgements by human listeners and automatic voice comparison. In *Proceedings of FONETIK 2010,* 63-69.

McDougall, K. (2013a). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, *20*(2), 163-172.

McDougall, K. (2013b). Earwitness evidence and the question of voice similarity. *British Academy Review* 21, 18-21.

McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016). The Speakers in the Wild (SITW) Speaker Recognition Database. In Proceedings of INTERSPEECH 2016, San Francisco, 818-822.

Nolan, F. (2003). A recent voice parade. *International Journal of Speech, Language and the Law*, *10*(2), 277-291.

Nolan, F. (2011). *Dynamic Variability in Speech: a Forensic Phonetic Study of British English, 2006-2007*. [data collection]. UK Data Service. SN: 6790, http://doi.org/10.5255/UKDA-SN-6790-1

Nolan, F., McDougall, K., & Hudson, T. (2011). Some Acoustic Correlates of Perceived (Dis)similarity between Same-accent Voices. In *ICPhS* (pp. 1506-1509).

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

Obin, N., & Roebel, A. (2016). Similarity search of acted voices for automatic voice casting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(9), 1642-1651.

Park, S. J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P. A., & Alwan, A. (2018). Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles. *The Journal of the Acoustical Society of America*, 144(1), 375-386.

Perrachione, T. K. (2019). Speaker recognition across languages. In: Frühholz, S., & Belin, P. (Eds.). *The Oxford Handbook of Voice Perception*, Oxford: Oxford University Press. 515-538.

Perrachione, T. K., Furbeck, K. T., & Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, *146*(5), 3384–3399.

Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, 10, 237–248

Remez, R. E., Fellowes, J. M., & Nagel, D. S. (2007). On the perception of similarity among talkers. *The Journal of the Acoustical Society of America*, *122*(6), 3688-3696.

Rietveld, A. C. M., & Broeders, A. P. A. (1991). Testing the fairness of voice identity parades: the similarity criterion. In *Proceedings of the 12th International Congress of Phonetic Sciences*, Aix-en-Provence. (Vol. 5, pp. 46-49).

San Segundo, E., Foulkes, P., & Hughes, V. (2016). Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli. In *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology, Parramatta, Australia.* 309-312.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Multidimensional Scaling: Theory, Methods, and Applications*. Academic Press.

Sherrin, C. (2015). Earwitness evidence: the reliability of voice identifications. *Osgoode Hall LJ*, *52*, 819-862.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research*, *21*(2), 265-275.