How do ASR systems perceive voice similarity?

Linda Gerlach¹, ², Kirsty McDougall¹, Finnian Kelly², Anil Alexander²
¹University of Cambridge, ²Oxford Wave Research
{lg589|kem37@cam.ac.uk, linda|finnian|anil@oxfordwaveresearch.com}

IAFPA conference Marburg (online), 22.-25.08.2021
Can we measure how similar these voices are?
Selection of a set of voices (foils) judged to lie within an appropriate range of similarity regarding “accent, inflection, pitch, tone and speed of the speech used” (Home Office 2003: point 15)

Automating the foil selection (under supervision of a forensic expert) could

- Allow for a larger pool of potential foils to be considered
- Provide a more objective selection process
- Increase speed and reduce costs

Problem:
The preparation of a voice parade is highly time consuming, costly, and subjective.
Further applications

- **Voice parades**
  foils with similar voices to the suspect

- **Voice banking**
  Original voice
  Synthesised voice

- **Voice casting**
  Voice actor

- **Selection of relevant populations for forensic speaker recognition**

© OxfordWaveResearch

© University of Cambridge
Gerlach et al. (2020):

Could perhaps an automatic speaker recognition (ASR) system relying on perceptually relevant features be used to select similar-sounding speakers for voice parades?

Small-scale pilot study assessing the similarity of 10 SSBE (Standard Southern British English) speakers

Correlation of lay-listener ratings of perceived voice similarity and ASR scores (auto-phonetic features + F0, i-vector)

Results showed a statistically significant, positive correlation
Research questions

- Is there a correlation between listener ratings and automatic estimates of perceived voice similarity within different groups of speakers with the same English accent and, if so, how does it vary?

- Is there a such a correlation for speakers with a different accent of English and, if so, how does it compare?

- How do different ASR feature extraction and speaker modelling approaches compare in approximating ratings of perceived voice similarity?
Speaker databases & groups

- **DyViS** (Nolan et al. 2009)
  - Database: 100 male SSBE speakers, 18-25 y. o.
  - Experiment sets: 3 x 15 speakers (D1, D2, D3)

- **YorViS** (McDougall et al. 2015)
  - Database: 20 male speakers of York English, 18-25 y. o.
  - Experiment sets: 1 x 15 speakers (Y)

- **WYRED** (Gold et al. 2018)
  - Database: 180 male speakers of West Yorkshire Englishes from Bradford, Kirklees, and Wakefield (60 per location), 18-30 y. o.
  - Experiment sets: 1 x 15 Bradford speakers (W1), 1 x 15 Wakefield speakers (W2)

From these databases, only Task 1 and 2 recordings containing spontaneous speech in studio quality were used.
Listener experiment

- Collection of listener ratings as part of the VoiceSim and IVIP (‘Improving Voice Identification Parades’) projects, made available for this study

- Paired speaker comparisons for each speaker group (D1, D2, D3, Y, W1, W2)
- Per speaker group 120 comparisons (incl. same-speaker comparisons)
- Two speech samples per speaker from Task 2 of the databases, ~3 s duration

- Voice similarity ratings from 120 listeners
  - 18-40 y. o., English L1, no hearing impairments, roughly balanced for sex
  - 20 listeners per speaker group
  - Similarity judged on a scale from 1 – very similar to 9 – very different
Automatic experiment

- **VOCALISE** forensic automatic speaker recognition software (Alexander et al. 2016, Kelly et al. 2019)

- **Comparison of results from five pre-trained sessions** combining different feature extraction methods and speaker modelling approaches:
  - i-vector\_AP\_F0: i-vectors, auto-phonetic features (including F0, semitones of F0, first derivatives, F1 to F4)
  - i-vector\_AP: i-vectors, auto-phonetic features (F1 to F4, without F0)
  - i-vector\_MFCC: i-vectors, spectral features
  - x-vector\_MFCC: x-vectors, spectral features
  - x-vector\_AP: x-vectors, auto-phonetic features (F1 to F4, without F0)

- Probabilistic Linear Discriminant Analysis (PLDA) to obtain comparison scores
Background: Automatic speaker recognition

What are i-vectors and x-vectors?
- Successors of Gaussian Mixture Models (GMMs) and Adapted Gaussian Mixture Models (GMM-UBM) (Reynolds & Rose 1995, Reynolds et al. 2000)
- Compact representations of a speaker with a fixed number of dimensions that can be directly compared resulting in a score

What is the difference?
- i-vectors are based on Factor Analysis (Dehak et al. 2011)
- x-vectors are based on Deep Neural Networks (Snyder et al. 2018)
Comparison of all speakers within each of the six speaker groups
Comparison scores used as voice similarity estimates
Two samples per speaker (~4 min each)
Normalisation of comparison scores using Bio-Metrics to ensure comparability across speaker groups
Averaging of comparison scores of same-speaker pairs to obtain one score per speaker pair
Evaluation and analysis

- Inversion of listener ratings so that $9 = \textit{very different}$ and $1 = \textit{very similar}$
- Calculation of 5% trimmed mean of the listener ratings for each comparison to adjust for bias
- Correlation analysis to explore the degree and direction of the relationship between listener ratings of voice similarity and automatically obtained scores

Analysis of different-speaker comparisons only

- Spearman rank correlation (2-tailed)
- Pearson’s correlation (2-tailed) to further assess the linearity of the relationship (for future analyses)
- For comparison of VOCALISE sessions: combination of all different-speaker scores and combination of the 5% trimmed mean listener ratings from all speaker groups (except YorViS Y due to miscalibration) before correlation
Results – same-accent background
(i-vector_AP_F0)

Calibrated VOCALISE score

Listener ratings (5% trimmed mean)

very different

very similar

D1 (same-speaker)

D1 (different-speaker)

D2 (same-speaker)

D2 (different-speaker)

D3 (same-speaker)

D3 (different-speaker)

Linear trend – D1 (different-speaker)

Linear trend – D2 (different-speaker)

Linear trend – D3 (different-speaker)
Results – same-accent background (i-vector_AP_F0)

-25 -20 -15 -10 -5 0 5 10

Calibrated VOCALISE score

Listener ratings (5% trimmed mean)

very different very similar

D1 (same-speaker)
D1 (different-speaker)
D2 (same-speaker)
D2 (different-speaker)
D3 (same-speaker)
D3 (different-speaker)

Linear trend – D1 (different-speaker)
Linear trend – D2 (different-speaker)
Linear trend – D3 (different-speaker)
Results – same-accent background (i-vector_AP_F0)

Calibrated VOCALISE score vs Listener ratings (5% trimmed mean)

- D1 (same-speaker)
- D1 (different-speaker)
- D2 (same-speaker)
- D2 (different-speaker)
- D3 (same-speaker)
- D3 (different-speaker)

Linear trend – D1 (different-speaker)
Linear trend – D2 (different-speaker)
Linear trend – D3 (different-speaker)
Results – different-accent background (i-vector_AP_F0)

<table>
<thead>
<tr>
<th>Speaker group</th>
<th>Spearman’s rho (2-tailed)</th>
<th>Pearson’s r ($r^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DyViS D1</td>
<td>0.484 ($p &lt; .001$)</td>
<td>0.503 (0.253)</td>
</tr>
<tr>
<td>DyViS D2</td>
<td>0.519 ($p &lt; .001$)</td>
<td>0.502 (0.252)</td>
</tr>
<tr>
<td>DyViS D3</td>
<td>0.357 ($p &lt; .001$)</td>
<td>0.366 (0.134)</td>
</tr>
<tr>
<td>WYRED W1</td>
<td>0.380 ($p &lt; .001$)</td>
<td>0.398 (0.158)</td>
</tr>
<tr>
<td>WYRED W2</td>
<td>0.551 ($p &lt; .001$)</td>
<td>0.631 (0.398)</td>
</tr>
<tr>
<td>YorViS Y</td>
<td>0.463 ($p &lt; .001$)</td>
<td><strong>0.704 (0.500)</strong></td>
</tr>
</tbody>
</table>

Influence of one very distinctive speaker
Results – VOCALISE sessions (combined speaker groups)

Spearman's rho

VOCALISE session

- x-vector_MFCC
- i-vector_AP
- i-vector_MFCC
- x-vector_AP
- i-vector_AP_F0
Findings

- Positive and statistically significant correlation results for the relationship between listener ratings and automatic estimates of perceived voice similarity within and across speaker groups of same and different accents.

- Variability of strength of correlation (and linearity) without apparent trends related to the different accents.

- Best performance observed with the VOCALISE session using auto-phonetic features including F0 and using i-vectors (i-vector_AP_F0).

- VOCALISE session with a potential to outperform i-vector_AP_F0: x-vector session using auto-phonetic features including F0.
Broader considerations

- Considerations must be given to:
  - the definition of the speaker search space based on the application
  - an application-dependent suitable degree of similarity between voices
Promising results indicating that perceived voice similarity can be assessed using an ASR approach

Future work:
- Explore an ASR approach that combines x-vectors with auto-phonetic features and F0
- Consider inter-rater agreement as point of reference for ‘sufficient’ correlation between listener ratings and automatic estimates
- Investigate score thresholds for perceived voice similarity for different applications
- Use lower quality audio
Special Thanks for Data

DyViS

VoiceSim

YorViS

WYRED

IVIP
References

- Reynolds, D. A., Rose, R. C., Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE trans. speech and audio processing, 3(1), 72-83, 1995.
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., X-vectors: Robust DNN Embeddings for Speaker Recognition, ICASSP 2018.
Questions?