A WYRED connection: x-vectors and forensic speech data

Anil Alexander¹, Finnian Kelly¹, and Erica Gold²

¹Oxford Wave Research Ltd, Oxford, United Kingdom
²University of Huddersfield, Huddersfield, United Kingdom

23 August 2021, IAFPA conference, Virtual Marburg
Introduction

• The West Yorkshire Regional English Database (WYRED; Gold et al. 2018) addressed the lacuna in a representative database of West Yorkshire English and includes short voicemail messages - not previously been tested for robustness and comparability in forensic speaker comparisons.

• Deep Neural Networks (DNNs) mark a new phase in the evolution of automatic speaker recognition technology, providing a powerful way to extract highly-discriminative speaker-specific features from a recording of speech.

• Rich meta-data, which is unique to the WYRED database allows for exploration of speaker profile information, including biological gender, age ranges, and spoken language.

• We examine the x-vector framework in VOCALISE and its performance on challenging comparisons within the forensically-relevant WYRED database.
The data

West Yorkshire Regional English Database (WYRED; Gold et al. 2018)

- Male
- Aged 18-30
- Grew up and went to school in Bradford, Kirklees or Wakefield
- From an English-only speaking household
- English = First & only language
- 60 Bradford, 60 Kirklees, and 60 Wakefield speakers (180 total)
## The tasks

### West Yorkshire Regional English Database (WYRED; Gold et al. 2018)

<table>
<thead>
<tr>
<th>Task and Number</th>
<th>Speaking Style</th>
<th>Interlocutor</th>
<th>Recording Quality</th>
<th>Length of Recording</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Police Interview</td>
<td>Formal, Spontaneous</td>
<td>Research Assistant 1</td>
<td>Studio</td>
<td>~ 20 mins.</td>
</tr>
<tr>
<td>2 Conversation with Accomplice</td>
<td>Less formal, Spontaneous</td>
<td>Research Assistant 2</td>
<td>Studio &amp; Telephone</td>
<td>~ 15 mins.</td>
</tr>
<tr>
<td>3 Paired Conversation</td>
<td>Casual, Spontaneous</td>
<td>Male participant from same area</td>
<td>Studio</td>
<td>~ 20 mins.</td>
</tr>
<tr>
<td>4 Voicemail Message</td>
<td>Time-pressured, Spontaneous</td>
<td>N/A</td>
<td>Studio &amp; Telephone+ Answerphone</td>
<td>~ 2 mins.</td>
</tr>
</tbody>
</table>
The tasks

West Yorkshire Regional English Database (WYRED; Gold et al. 2018)

<table>
<thead>
<tr>
<th>Task and Number</th>
<th>Speaking Style</th>
<th>Interlocutor</th>
<th>Recording Quality</th>
<th>Length of Recording</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Police Interview</td>
<td>Formal, Spontaneous</td>
<td>Research Assistant 1</td>
<td>Studio</td>
<td>~ 20 mins.</td>
</tr>
<tr>
<td>2 Conversation with Accomplice</td>
<td>Less formal, Spontaneous</td>
<td>Research Assistant 2</td>
<td>Studio &amp; Telephone</td>
<td>~ 15 mins.</td>
</tr>
<tr>
<td>3 Paired Conversation</td>
<td>Casual, Spontaneous</td>
<td>Male participant from same area</td>
<td>Studio</td>
<td>~ 20 mins.</td>
</tr>
<tr>
<td>4 Voicemail Message</td>
<td>Time-pressured, Spontaneous</td>
<td>N/A</td>
<td>Studio, Telephone+ Answerphone</td>
<td>~ 2 mins.</td>
</tr>
</tbody>
</table>
**WYRED Task 4**

- West Yorkshire Regional English Database (WYRED; Gold et al. 2018)
  - Participants left a (supposedly) urgent answerphone message in relation to the fictional crime from Tasks 1 and 2
  - Participant took on the role of a suspect in the fictional criminal investigation.
  - Answerphone message was to be left for their fictional brother
  - Participants provided with short prompts about incriminating evidence which they were to instruct their brother to hide or dispose of
  - **Purpose** - to elicit non-contemporaneous speech, recorded in the studio against a time constraint, and of telephone quality
A Cautionary Tale For Phonetic Analysis: The Variability of Speech Between and Within Recording Sessions [Ross, Earnshaw and Gold, ICPHS 2019 ]

i-vector based study with a different data subset from WYRED

Did not include Task 4 as the files were considered too short

Findings:

- Within-task results - perfect discrimination (0% EER)
- Within-session between-tasks – slightly increased EER
- Between-session – good performance (2-3% EER), but elevated compared to within-task

In this study, we decided to include Task 4 (shorter recordings), and to do a cross-session and cross-task comparisons using the next generation of algorithms.
An automatic speaker recognition pipeline

The technology has evolved, but the general pipeline has remained consistent
The i-vector and x-vector pipelines

speech → feature extraction → Deep Neural Network → UBM → i-vector

Deep Neural Network → x-vector
The x-vector pipeline

Deep Neural Network

features

High-dimensional, universal speaker space

Low-dimensional, speaker-specific space

The performance of x-vectors outperform that of i-vectors, particularly at short durations.

The x-vector DNN is capable of exploiting larger amounts of training data than the i-vector framework.
Task 4 Tel Vs Task 2 studio: example
Visualising speaker x-vectors in three dimensions

Speaker 60 stands out from the group – distinctive voice quality
Recordings were pre-processed using Praat TextGrid files to remove all audio content not pertaining to the speaker of interest.

All cross-task recording pairings (180 speakers) were compared using VOCALISE in x-vector PLDA mode.

We focus on the results involving Task 2 and Task 4 (answerphone message)
# Results: cross-task equal error rate (EERs)

<table>
<thead>
<tr>
<th></th>
<th>Task 1 studio</th>
<th>Task 2 studio</th>
<th>Task 3 studio</th>
<th>Task 2 tel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 4 studio</td>
<td>0.50%</td>
<td>1.06%</td>
<td>0.72%</td>
<td>2.50%</td>
</tr>
<tr>
<td>Task 4 tel+ans</td>
<td>1.07%</td>
<td>1.72%</td>
<td>1.47%</td>
<td>2.40%</td>
</tr>
</tbody>
</table>
Task 4 tel Vs Task 2 studio: Zoo plot
Speaker Profiling with x-vectors

Speaker profile information

Speaker x-vectors

LDA (Linear discriminant analysis)

Support Vector Machine (SVM) Classifier: optimally separate x-vectors from different classes based on a set of training data

Maximum margin separating x-vectors
Gender (biological sex) recognition

x-vector from an unknown speaker

Male x-vectors

Female x-vectors

Female label
Language recognition

Choose the language that classifies unknown speaker $x$-vector with the greatest margin

- Language 1
- Language 2
- Language 3

$x$-vector from an unknown speaker

Language 3 label
Age and height estimation

- x-vector from an unknown speaker
- Speaker Age A or Height H

Speaker age or height
Speaker profiling results

- Gender recognition: 99% correct in studio condition/ 98% in tel condition 😊
  - Speaker #93 is detected as female in both studio and tel (only person to be marked as female in both) and is the lightest in weight of ALL the participants
  - Speaker #93 Task 4 tel / Task 2 studio: 🎤 🎤

- Language recognition: 93% correct in studio condition 😊
  - Speaker 15 (Arabic) Task 2 studio: 🎤
  - Speaker 30 (German) Task 2 studio: 🎤

- Age estimation results are poor – narrow age range (mostly 20-25) and speaker demographic represents a large mismatch from the profiling training data 😞

- Height estimation was better but underestimated – also suffers from the training data mismatch, but generally gives higher estimates to taller speakers
Conclusions

- The controlled variability in the WYRED database enables structured speaker recognition evaluations of this kind, and the meta-data further supports the exploration of speaker profiling.

- The type of forensically-relevant recordings contained in Task 4 are not traditionally seen in databases collected for research.

- However, the results demonstrate that they are well-suited for automatic analysis with the DNN-based version of VOCALISE using x-vectors.

- The speaker profiling results with x-vectors demonstrate that:
  - Gender recognition generalises very well, producing very few errors.
  - Language recognition is good but challenged by UK regional accents.
  - Age and height estimation suffer from training/testing data mismatch – in-domain training or adaptation could improve performance.
References


