Detecting emotional and cognitive changes using privacy preserving features in spontaneous interaction speech data

Saturnino Luz
Usher Institute
Edinburgh Medical School

19 November 2019
WHINN’19, Odense, Denmark

Supporting Active Ageing through multimodal coaching

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661
The sensor era and it’s promises

- Ambient intelligence, mobile and wearable devices, a multitude of sensors, companion robots, ...
- Monitor health by analysing everyday (inter)activity data
- Applications in elderly care, and coaching for healthy ageing

My lab’s focus is on monitoring cognitive function and affect through speech analysis.
The sensor era and it’s promises (and threats)

- Ambient intelligence, mobile and wearable devices, a multitude of sensors, companion robots, ...
- Monitor health by analysing everyday (inter)activity data
- Applications in elderly care, and coaching for healthy ageing

My lab’s focus is on monitoring cognitive function and affect through speech analysis.
Examples of recent research (in the context of SAAM)

- Use of speech for
  - detection of mental health problems
  - detection of cognitive changes
  - monitoring of emotional states

- Beyond SAAM
  - collection detailed multimodal (speech, 3d video, visuospatial data) longitudinal datasets (genetic, medical, family history and lifestyle information)
  - characterisation and early detection of Alzheimer’s dementia
  - interactive technology for people living with AD and their carers (e.g. speech therapy)
The SAAM setup

The hardware:
- Matrix Creator board:
  - a microphone array,
  - an inertial measurement unit,
  - and several other sensors
- mounted on a Raspberry Pi 3 B+
- Alternatively, sensors and depth camera mounted on an Intel Up! Board
Speech (and social signal) data protection has not received as much attention as other modalities (e.g. text and structured records).

Privacy protection and preservation for speech can be considered from different perspectives:

➤ protection of a person’s identity (Id),
➤ protection of the spoken content (C), and
➤ protection of inferences (In) one may be able to draw from characteristics of a person’s voice (such as cognitive or emotional status)
Guarding against privacy leaks usually causes ML inference utility loss\(^1\)

<table>
<thead>
<tr>
<th>Privacy leak</th>
<th>Threat dimension</th>
<th>Type of utility loss</th>
<th>training</th>
<th>inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>text leak</td>
<td>C, Id</td>
<td>voice diversity</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>voice attribute leak</td>
<td>In</td>
<td>text authenticity</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>voiceprint leak</td>
<td>Id, In</td>
<td>speech quality</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>membership leak</td>
<td>In</td>
<td>data use clarity</td>
<td>+++</td>
<td>+</td>
</tr>
</tbody>
</table>

A taxonomy of privacy preserving ML

Privacy-preserving training
- Parameter transmission
- Distributed machine learning
- Anonymization
- Data synthesis
- Cryptography
- Data obfuscation
- Multiparty computation
- CryptoNets
- ObfNet

Privacy-preserving testing
- Model personalization
- Conventional distributed ML
- Federated Learning
- Vocalization graphs, acoustic features
- Homomorphic encryption
- Multiplicative perturbation
- Additive perturbation
- Generative obfuscation


This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661.
A taxonomy of privacy preserving ML

Privacy-preserving training
- Parameter transmission
  - Model personalization
  - Conventional distributed ML
  - Federated Learning

Privacy-preserving testing
- Data transmission
  - Anonymization
    - Data synthesis
      - Cryptography
      - Data obfuscation
      - Multiparty computation
      - CryptoNets
      - ObfNet
      - Homomorphic encryption
      - Multiplicative perturbation
      - Additive perturbation
      - Generative obfuscation

Vocalization graphs, acoustic features

Simple anonymisation by feature extraction

Live Audio Stream → Voice Activity Detection → Saving audio clip → Watchdog → Feature extraction → Speaker recognition → Saving features → Deleting audio clip → Behaviour Analysis

- Affective and emotional processing
- Cognitive decline detection

Example: ATD diagnosis, early detection

- **Need for tests** that can detect Mild Cognitive Impairment (MCI) and Alzheimer’s type Dementia (ATD)
  - characterise impairment
  - monitor interventions/therapy
- **Need for methods** (other than, but correlated to, neuroimaging) to monitor cognitive status and detect ATD early on:
  - pre-clinical stage, for secondary prevention, or
  - earlier still, for research
- If such methods exist, they should be catching subtle cognitive changes early on

Focus on speech and language

- Much information on cognitive status can be gathered through speech
- Data sources:
  - word tests,
  - narration (scene descriptions),
  - interviews,
  - spontaneous conversations, ...
- Existing Datasets:
  - DementiaBank, the Pitt Dataset,
  - Carolina conversations corpus,
  - ....
The Pitt Dataset from DementiaBank

Recorded speech data for a number of neuropsychological tests:

- Fluency
- Word recall
- Sentence production
- Cambridge Cookie Theft test:
  - Probable AD speech
  - Normal control speech

Automatic detection of AD speech

▶ Data, a subset of the Pitt dataset (cookie test):

<table>
<thead>
<tr>
<th>Control</th>
<th>MCI</th>
<th>Memory</th>
<th>PossibleAD</th>
<th>ProbableAD</th>
<th>Vascular</th>
</tr>
</thead>
<tbody>
<tr>
<td>242</td>
<td>43</td>
<td>3</td>
<td>21</td>
<td>236</td>
<td>5</td>
</tr>
</tbody>
</table>

▶ Matched by ages and gender:

<table>
<thead>
<tr>
<th>Age interval</th>
<th>Control</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(45, 50]</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(50, 55]</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>(55, 60]</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>(60, 65]</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>(65, 70]</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>(70, 75]</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>(75, 80]</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
A pilot study of content-free features

Audio Processing

Feature extraction

Classification

Model training

\begin{align*}
P(a|F) & \propto P(F_1 = v_1, \ldots, F_n = v_n|a) \\
& = \prod_{i=1}^{n} P(F_i = v_i|a)
\end{align*}
Results (Pitt dataset)

- Overall accuracy of 68%, (up to 77% on improved dataset and LDA; work in progress)
- $F_1$ scores of 70% for the control class and 64% for the ATD class
- (“gold standard” = diagnostic established through a combination of neuropsychological and neurologic tests)
A more detailed study on the Pitt monologues

Fully balanced, acoustically enhanced dataset

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Male AD</th>
<th>Female AD</th>
<th>Male non-AD</th>
<th>Female non-AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>[50, 55)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>[55, 60)</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>[60, 65)</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>[65, 70)</td>
<td>10</td>
<td>14</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>[70, 75)</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>[75, 80)</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>46</td>
<td>36</td>
<td>46</td>
</tr>
</tbody>
</table>

Assessed several voice feature sets:

- **emobase**: MFCC, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features with their first and second order derivatives, and functionals. (988 features)

- **ComParE**: energy, spectral, MFCC, and voicing related low-level descriptors, including harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness, and functionals. (6,373 features)

- **eGeMAPS**: basic set of theoretically motivated acoustic features. (88 features)

- **MRCG functionals**: a cochleagram is generated by applying the gammatone filter to the audio signal, decomposing it in the frequency domain so as to mimic the human auditory filters. MRCG uses the time-frequency representation to encode the multi-resolution power distribution of the audio signal. (6,912 features; 768 per 20ms frame: 256 MRCG 256 ∆ MRCG and 256 ∆∆ MRCG)
Active data representation (ADR)

ADR feature extraction process:
1. Audio segmentation
2. Clustering
3. Generation of histograms for segment duration and number
4. Computation of rate of change in cluster membership
5. Normalisation (L1 norm)

Supporting Active Ageing through multimodal coaching

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661

# Classification results

Majority vote classification using full feature sets.

<table>
<thead>
<tr>
<th>Features</th>
<th>LDA</th>
<th>DT</th>
<th>SVM</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>emobase</td>
<td>53.66</td>
<td>56.10</td>
<td>56.71</td>
<td>57.93</td>
</tr>
<tr>
<td>ComParE</td>
<td><strong>61.59</strong></td>
<td>46.95</td>
<td>54.88</td>
<td>58.54</td>
</tr>
<tr>
<td>eGeMAPS</td>
<td>50.61</td>
<td>51.83</td>
<td>50.61</td>
<td>60.98</td>
</tr>
<tr>
<td>MRCG</td>
<td>50.61</td>
<td>54.88</td>
<td>56.10</td>
<td>56.10</td>
</tr>
</tbody>
</table>

## ADR classification results

<table>
<thead>
<tr>
<th>Features</th>
<th>LDA, m</th>
<th>DT, m</th>
<th>1NN, m</th>
<th>SVM, m</th>
<th>RF, m</th>
</tr>
</thead>
<tbody>
<tr>
<td>emobase</td>
<td>56.10, 30</td>
<td>66.46, 20</td>
<td>54.88, 80</td>
<td>45.12, 15</td>
<td>60.98, 25</td>
</tr>
<tr>
<td>ComParE</td>
<td>57.93, 35</td>
<td>68.90, 95</td>
<td>55.49, 100</td>
<td>59.76, 35</td>
<td>60.37, 95</td>
</tr>
<tr>
<td>eGeMAPS</td>
<td><strong>77.44</strong>, 85</td>
<td>71.34, 30</td>
<td>54.27, 65</td>
<td>52.44, 20</td>
<td>71.34, 30</td>
</tr>
<tr>
<td>MRCG</td>
<td>59.76, 5</td>
<td>69.51, 15</td>
<td>52.44, 95</td>
<td>59.76, 15</td>
<td>63.41, 15</td>
</tr>
<tr>
<td>mean</td>
<td>62.81</td>
<td><strong>69.05</strong></td>
<td>54.27</td>
<td>54.27</td>
<td>64.03</td>
</tr>
</tbody>
</table>
Combining feature sets

Analysing the performance of the ADR for different feature sets with a DT classifier:

As different ADRs seem to capture different aspects of classification, “ADR fusion” should improve results:
Comparisons with the state-of-the-art

<table>
<thead>
<tr>
<th>Study</th>
<th>Accuracy</th>
<th>Modality</th>
<th>Fully Automatic</th>
<th>Privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our approach</td>
<td>78.7%</td>
<td>acoustic</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Hernández et al., 2018</td>
<td>62.0%</td>
<td>acoustic</td>
<td>yes</td>
<td>no(?)</td>
</tr>
<tr>
<td>Luz, 2017</td>
<td>68.0%</td>
<td>acoustic</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Mirheidari et al., 2018</td>
<td>62.3%</td>
<td>text</td>
<td>yes (ASR)</td>
<td>no</td>
</tr>
<tr>
<td>Fraser et al., 2016</td>
<td>81.9%</td>
<td>text/acoustic</td>
<td>no</td>
<td>no (text)</td>
</tr>
<tr>
<td>Yancheva &amp; Rudzicz, 2016</td>
<td>80.0%</td>
<td>text/acoustic</td>
<td>no (text)</td>
<td>no</td>
</tr>
<tr>
<td>Hernández et al., 2018</td>
<td>68.0%</td>
<td>text</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Mirheidari et al., 2018</td>
<td>75.6%</td>
<td>text</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
Detecting cognitive decline through dialogue

▶ **Hypothesis:** AD patient dialogues show identifiable patterns during dialogue interactions, such as disrupted turn taking and differences in speech rate.

▶ **Spontaneous, conversational speech** from the Carolina Conversations Collection (CCC), a subset
  ▶ **21 patients with a diagnosis of Alzheimer’s disease** (15 females, 6 males),
  ▶ and **17 patients (12 females, 5 males)** with other diseases (excluding neuropsychological conditions)
  ▶ **matching age range and gender frequencies.**

▶ **Representation:**
  ▶ **vocalisation graphs (VGO)**
  ▶ **vocalisation graphs + speech rate (VGS)**
Results for additive logistic regression models

### AD detection results for the VGO data representation scheme.

<table>
<thead>
<tr>
<th></th>
<th>AD</th>
<th>non-AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.812</td>
<td>0.714</td>
</tr>
<tr>
<td>Precision</td>
<td>0.765</td>
<td>0.769</td>
</tr>
<tr>
<td>Recall</td>
<td>0.812</td>
<td>0.714</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.788</td>
<td>0.741</td>
</tr>
<tr>
<td>Precision</td>
<td>0.667</td>
<td>0.792</td>
</tr>
<tr>
<td>Recall</td>
<td>0.722</td>
<td>0.729</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.685</td>
<td>0.721</td>
</tr>
</tbody>
</table>

Overall accuracy (LOOCV): 0.811

### Results for the VGS data representation scheme.

<table>
<thead>
<tr>
<th></th>
<th>AD</th>
<th>non-AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.882</td>
<td>0.769</td>
</tr>
<tr>
<td>Precision</td>
<td>0.833</td>
<td>0.833</td>
</tr>
<tr>
<td>Recall</td>
<td>0.882</td>
<td>0.769</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.857</td>
<td>0.800</td>
</tr>
<tr>
<td>Precision</td>
<td>0.796</td>
<td>0.708</td>
</tr>
<tr>
<td>Recall</td>
<td>0.833</td>
<td>0.708</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.811</td>
<td>0.700</td>
</tr>
</tbody>
</table>

Overall accuracy (LOOCV): 0.865
Comparing models

Compared accuracy results obtained with different classification algorithms, on VGS-based datasets.

<table>
<thead>
<tr>
<th>Classification method</th>
<th>Accuracy (LOOCV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>75.7%</td>
</tr>
<tr>
<td>Real Adaboost</td>
<td>86.5%</td>
</tr>
<tr>
<td>Decision trees</td>
<td>86.5%</td>
</tr>
<tr>
<td>SVM</td>
<td>83.7%</td>
</tr>
<tr>
<td>Random forests</td>
<td>81.1%</td>
</tr>
</tbody>
</table>
Remaining Challenges

Feature extraction anonymisation:
- Does not seem to hinder ML utility
- And is suitable for IoT devices,
- But is vulnerable to:
  - Background knowledge attacks,
  - Linkage attacks, and
  - membership leaks
- Further research on lightweight encryption and privacy-preserving ML inference for speech data is needed.
References


