Comparing detection methods for pause-internal particles (PINTs)

Mikey Elmers
Department Language Science and Technology at Saarland University
elmers@lst.uni-saarland.de
Introduction

- Silent segments
- Breath noises
  - Inhalations
  - Exhalations
- Filler particles
  - „äh“ and „ähm“ in German
  - „uh“ and „uhm“ in English
- Tongue clicks
PINTs TTS

• Silent segments improve digit recollection (Elmers et al. 2021a)
• Breath noises improve sentence recollection (Elmers et al. 2021b)
• Filler particles improve TTS by reducing cognitive load for listener (Dall et al. 2016)
• Quality of training data is important for TTS applications (Henter et al. 2016)
Silent Segment
Breath Noises
Filler Particles
Clicks
Co-Ocurrence

![Co-Occurrence Diagram]

- kHz
- Time (s)
- sil
- csil
- inh
- sil
- sp
- uh
- sp
- uh
- exh
Co-Occurrence

• Modeling multiple PINTs improved classification accuracy of surrounding non-verbal vocalizations (Condron et al. 2021)

• PINTs are usually:
  ▪ Condensed to “other” class
  ▪ Ignored altogether
Aim

• Implement state-of-the-art methods for detecting PINTs
• Classification of PINTs in German
• Classify PINTs using three models:
  ▪ General neural network (NN)
  ▪ Convolutional neural network (CNN)
  ▪ Recurrent neural network (RNN)
• Hypotheses:
  ▪ RNN will outperform other models
  ▪ Simultaneous modeling improves PINTs classification
Methods

• Corpus Information:
  ▪ Pool Corpus (Jessen et al. 2005)
  ▪ 100 males (21-63 years old; mean age 39 years old)
  ▪ Native speakers of German
  ▪ Spontaneous speech task (i.e. picture description task)
  ▪ Similar to board game Taboo
Methods

• Annotations:
  ▪ 100 files (124-374 s; mean dur 223 s; total dur 6.2 hours)
  ▪ Sampled at 16 kHz
  ▪ 17,641 annotated PINTs
    • Silent segments, inhalations, exhalations, two types of filler particles („uh“ and „uhm“), and clicks
  ▪ Other PINTs and disfluencies were excluded due to their infrequent occurrence
Methods

- Annotated PINTs overview
  - Min, max, mean, and sd measured in seconds
  - Total measured in minutes

<table>
<thead>
<tr>
<th>class</th>
<th>count</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
<th>total</th>
<th>prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>silent segment</td>
<td>10,237</td>
<td>0.01</td>
<td>20.01</td>
<td>0.65</td>
<td>0.95</td>
<td>111.04</td>
<td>29.92%</td>
</tr>
<tr>
<td>inhalation</td>
<td>2,891</td>
<td>0.05</td>
<td>2.10</td>
<td>0.51</td>
<td>0.27</td>
<td>24.79</td>
<td>6.68%</td>
</tr>
<tr>
<td>exhalation</td>
<td>1,887</td>
<td>0.03</td>
<td>3.23</td>
<td>0.38</td>
<td>0.28</td>
<td>12.15</td>
<td>3.27%</td>
</tr>
<tr>
<td>filler (uh)</td>
<td>1,156</td>
<td>0.04</td>
<td>1.44</td>
<td>0.35</td>
<td>0.16</td>
<td>6.81</td>
<td>1.83%</td>
</tr>
<tr>
<td>filler (uhm)</td>
<td>549</td>
<td>0.15</td>
<td>2.64</td>
<td>0.53</td>
<td>0.25</td>
<td>4.85</td>
<td>1.30%</td>
</tr>
<tr>
<td>click</td>
<td>921</td>
<td>0.00</td>
<td>0.50</td>
<td>0.06</td>
<td>0.05</td>
<td>0.96</td>
<td>0.25%</td>
</tr>
</tbody>
</table>
Methods

• Data pre-processing:
  - 13 mel-frequency cepstral coefficients (MFCCs)
  - Frame size 93 ms
  - Hop length 23 ms
  - Zero-padding
Methods

• Data pre-processing:
  ▪ Models trained on nine classes
    • Silent segments
    • Inhalation
    • Exhalation
    • Two FPs (“uh” and “uhm”)
    • Clicks
    • Speech
    • Task change
    • Zero-padding
Methods

• Model Information:
  ▪ Same hyperparameters
  ▪ Similar number of layers
  ▪ Same number of neurons for those layers
  ▪ Sparse categorical cross entropy loss function
  ▪ Learning rate of 0.0001
  ▪ Adam optimizer
  ▪ Batch size of 32
  ▪ Trained for 40 epochs
Methods – Neural Network

1. MFCC Extraction

2. Neural Network

- Flatten
- 30% dropout
- ReLU
- Softmax
- Dense
- 64
- 64
- 9
- Output probability
Methods – Convolutional Neural Network

1. MFCC Extraction
2. Convolutional Neural Network

Output probability

- Flatten
- 1D max pooling
- 1D Conv
- 30% dropout
- ReLU
- Softmax
- Dense
- Batch normalization
Methods – Recurrent Neural Network

1. MFCC Extraction
2. Recurrent Neural Network

Output probability

- LSTM
- 30% dropout
- ReLU
- Softmax
- Dense
## Results

<table>
<thead>
<tr>
<th>Class</th>
<th>NN</th>
<th>CNN</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>silent segment (sil)</td>
<td>64,971</td>
<td>66,494</td>
<td>64,771</td>
</tr>
<tr>
<td>inhalation</td>
<td>4,141</td>
<td>5,111</td>
<td>4,214</td>
</tr>
<tr>
<td>exhalation</td>
<td>3,215</td>
<td>3,173</td>
<td>2,812</td>
</tr>
<tr>
<td>filler (uh)</td>
<td>60</td>
<td>53</td>
<td>38</td>
</tr>
<tr>
<td>filler (uhm)</td>
<td>68</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>click</td>
<td>209</td>
<td>181</td>
<td>165</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>72,664</td>
<td>75,092</td>
<td>72,050</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>sil</th>
<th>inh</th>
<th>exh</th>
<th>uh</th>
<th>uhm</th>
<th>click</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>silent segment (sil)</td>
<td>64,971</td>
<td>2,743</td>
<td>789</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>68,503</td>
</tr>
<tr>
<td>inhalation</td>
<td>4,141</td>
<td>10,372</td>
<td>58</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14,571</td>
</tr>
<tr>
<td>exhalation</td>
<td>3,215</td>
<td>497</td>
<td>2,188</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5,900</td>
</tr>
<tr>
<td>filler (uh)</td>
<td>60</td>
<td>3</td>
<td>34</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>97</td>
</tr>
<tr>
<td>filler (uhm)</td>
<td>68</td>
<td>4</td>
<td>33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>105</td>
</tr>
<tr>
<td>click</td>
<td>209</td>
<td>85</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>301</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>72,664</td>
<td>13,704</td>
<td>3,108</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>89,477</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>sil</th>
<th>inh</th>
<th>exh</th>
<th>uh</th>
<th>uhm</th>
<th>click</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>silent segment (sil)</td>
<td>66,494</td>
<td>1,375</td>
<td>754</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>68,624</td>
</tr>
<tr>
<td>inhalation</td>
<td>5,111</td>
<td>9,351</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14,562</td>
</tr>
<tr>
<td>exhalation</td>
<td>3,173</td>
<td>336</td>
<td>2,532</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6,041</td>
</tr>
<tr>
<td>filler (uh)</td>
<td>53</td>
<td>2</td>
<td>27</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>82</td>
</tr>
<tr>
<td>filler (uhm)</td>
<td>80</td>
<td>5</td>
<td>20</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>116</td>
</tr>
<tr>
<td>click</td>
<td>181</td>
<td>73</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>265</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>75,092</td>
<td>11,142</td>
<td>3,444</td>
<td>-</td>
<td>11</td>
<td>1</td>
<td>89,690</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>sil</th>
<th>inh</th>
<th>exh</th>
<th>uh</th>
<th>uhm</th>
<th>click</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>silent segment (sil)</td>
<td>64,771</td>
<td>1,813</td>
<td>811</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>67,395</td>
</tr>
<tr>
<td>inhalation</td>
<td>4,214</td>
<td>10,098</td>
<td>113</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14,425</td>
</tr>
<tr>
<td>exhalation</td>
<td>2,812</td>
<td>394</td>
<td>2,308</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5,514</td>
</tr>
<tr>
<td>filler (uh)</td>
<td>38</td>
<td>2</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>53</td>
</tr>
<tr>
<td>filler (uhm)</td>
<td>50</td>
<td>2</td>
<td>17</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>72</td>
</tr>
<tr>
<td>click</td>
<td>165</td>
<td>74</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>250</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>72,050</td>
<td>12,383</td>
<td>3,270</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>87,709</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>85.6%</td>
<td>53.5%</td>
<td>41.6%</td>
<td>40.5%</td>
</tr>
<tr>
<td>CNN</td>
<td>86.1%</td>
<td>53.2%</td>
<td>41.9%</td>
<td>41.8%</td>
</tr>
<tr>
<td>RNN</td>
<td>86.1%</td>
<td>69.0%</td>
<td>42.1%</td>
<td>41.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>sil</th>
<th>inh</th>
<th>exh</th>
<th>uh</th>
<th>uhm</th>
<th>click</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>94.8%</td>
<td>71.2%</td>
<td>31.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.3%</td>
</tr>
<tr>
<td>CNN</td>
<td>96.9%</td>
<td>64.2%</td>
<td>41.9%</td>
<td>0.0%</td>
<td>9.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>RNN</td>
<td>96.1%</td>
<td>70.0%</td>
<td>41.9%</td>
<td>0.0%</td>
<td>4.2%</td>
<td>1.2%</td>
</tr>
</tbody>
</table>
Conclusions

• All models performed similarly

• Hypotheses:
  ▪ 1) RNN should perform best since it considers temporal information
    • RNN did not perform much better than NN or CNN
Conclusions

• Hypotheses:
  ▪ 2) Simultaneous modeling can improve classification accuracy of surrounding PINTs
    • Simultaneous modeling didn’t improve accuracy for surrounding PINTs
    • All models unable to classify FPs and clicks
    • FPs too close to speech category
    • Clicks often misclassified as silent segments
      – short duration
      – drawback of only using MFCCs as input
Conclusions

• Model classified:
  ▪ Silent segments very well
  ▪ Inhalations well
  ▪ Exhalations with middling success

• Accurate PINTs classification dependent on:
  ▪ Annotation quality
  ▪ Annotation quantity
  ▪ Models started with high accuracy and improved minimally
Conclusions

• Improvement to PINTs detection:
  ▪ Increase number of occurrences
  ▪ Especially for infrequent PINTs

• Future work
  ▪ Investigate other acoustic features
  ▪ Train using spectrogram images
  ▪ Implement PINTs classification into TTS pipeline


Thank you!

http://pauseparticles.org/
Conclusions

• Model classified:
  ▪ Silent segments very well
  ▪ Inhalations well
  ▪ Exhalations with middling success

• Accurate PINTs classification dependent on:
  ▪ Annotation quality
  ▪ Annotation quantity
  ▪ Models started with high accuracy and improved minimally