High Level Analysis of the Psychotherapy Sessions

Alexander Polok  
Pavel Matějka

Faculty of Information Technology, Brno University of Technology

Design a system, DeePsy demonstrator, which shows the analysis of the progress of the psychotherapy meetings and does provide:

- a systematic feedback on therapeutic work,
- a sophisticated system of questionnaires,
- an automatic analysis of session content using deep learning.

The dataset has been partially annotated within this thesis. This dataset consists of 11 online, five sessions recorded on a mobile phone and 32 psychotherapy sessions. To prepare speech encoder, 700 hours of collected unannotated psychotherapy sessions, referred as DeePsyUnsupervised, were used. Downstream speech/text models were finetuned on multi-domain corpus displayed in Figure 3.

**Voice Activity Detection – VAD**

The first, but very crucial, step for building a system for extracting complex entities is to extract speech from the recordings. This is done using a voice activity detection system - VAD. The originally integrated voice system vad baseline, based on a two-layer neural network [6] was too aggressive, thus experiments with different architectures summarized in Table 1 were conducted.

**Automatic Speech Recognition – ASR**

Classifying or extracting more complex features from dialogues requires high-quality speech and text features. However, the automatic speech transcription itself, which is based on the hybrid architecture CNN-TDNN-HMM supplemented with an n-gram language model, achieved a relatively high error rate of 28.10% WER.

Because of this, experiments were conducted with models based on the Transformer architecture, as described in the thesis. Major steps to obtain those results are displayed in Table 2.

- **Motivation**
- **Proposed System**
- **The Data**
- **Automatic Speaker Recognition – ASR**
- **Inference**
- **Conclusions**
- **References**

---

**Table 1.** The error rate of pretrained and finetuned (adapted by horizontal link) voice activity detection systems was evaluated on the DeePsyTest dataset using the following metrics: Detection Error Rate – DER, False Alarm – FA, and Miss Rate – M.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DER [%]</td>
<td>10.63</td>
<td>20.05</td>
<td>9.92</td>
<td>12.38</td>
<td>14.46</td>
<td>10.63</td>
</tr>
<tr>
<td>FA [%]</td>
<td>9.35</td>
<td>13.81</td>
<td>10.03</td>
<td>3.91</td>
<td>10.91</td>
<td>7.22</td>
</tr>
<tr>
<td>M [%]</td>
<td>1.29</td>
<td>16.84</td>
<td>7.90</td>
<td>9.08</td>
<td>3.56</td>
<td>1.16</td>
</tr>
</tbody>
</table>

**Table 2.** Following the steps described in this thesis, a system based on the FERNET [5] model was finetuned to obtain results visible in this table.

<table>
<thead>
<tr>
<th>Task</th>
<th>System</th>
<th>Metric</th>
<th>Value</th>
<th>Rel. imp. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice activity detection</td>
<td>PyanNet</td>
<td>WER [%]</td>
<td>2.99</td>
<td>+49.06</td>
</tr>
<tr>
<td>Diarization</td>
<td>Adapted VBx [4]</td>
<td>DER [%]</td>
<td>6.10</td>
<td>-1.16</td>
</tr>
<tr>
<td>Sentiment classification</td>
<td>CZERT [7]</td>
<td>macro F1</td>
<td>0.45</td>
<td>-</td>
</tr>
<tr>
<td>Therapeutic intervention classification</td>
<td>FERNET</td>
<td>macro F1</td>
<td>0.47/0.69</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 3.** Summary table of the best results in the respective tasks.

---

**References**