High Level Analysis of the Psychotherapy Sessions

Alexander Polok Pavel Matějka

Faculty of Information Technology, Brno University of Technology

Motivation

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.

Figure 1. Illustration of the treatment process between a client with mental disorder and a therapist, adapted from Freepik.com designed by *pch.vector*.

Proposed System

A schematic diagram illustrating the processing of the recording within the system is shown in Figure 2. First, **Voice Activity Detection** (VAD) is performed, triggering **Automatic Speech Recognition** (ASR), over the active sections, followed by **diarization**.

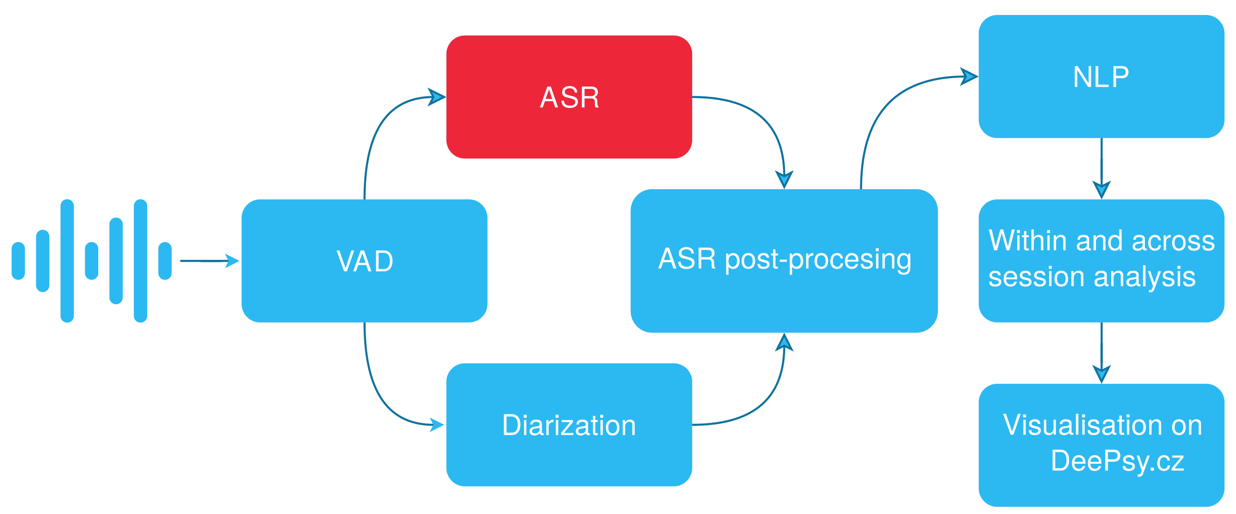

Figure 2. Schematic diagram of DeePsy session recording processing.

The outputs of these systems are then combined, and punctuation is added. Subsequently, **sentiment analysis, therapist intervention, and verb tense classifications** are processed. Speech and text features are then analyzed within a session and between sessions.

The Data

The presented experiments in the following section were evaluated on the **DeeePsyTest dataset**.

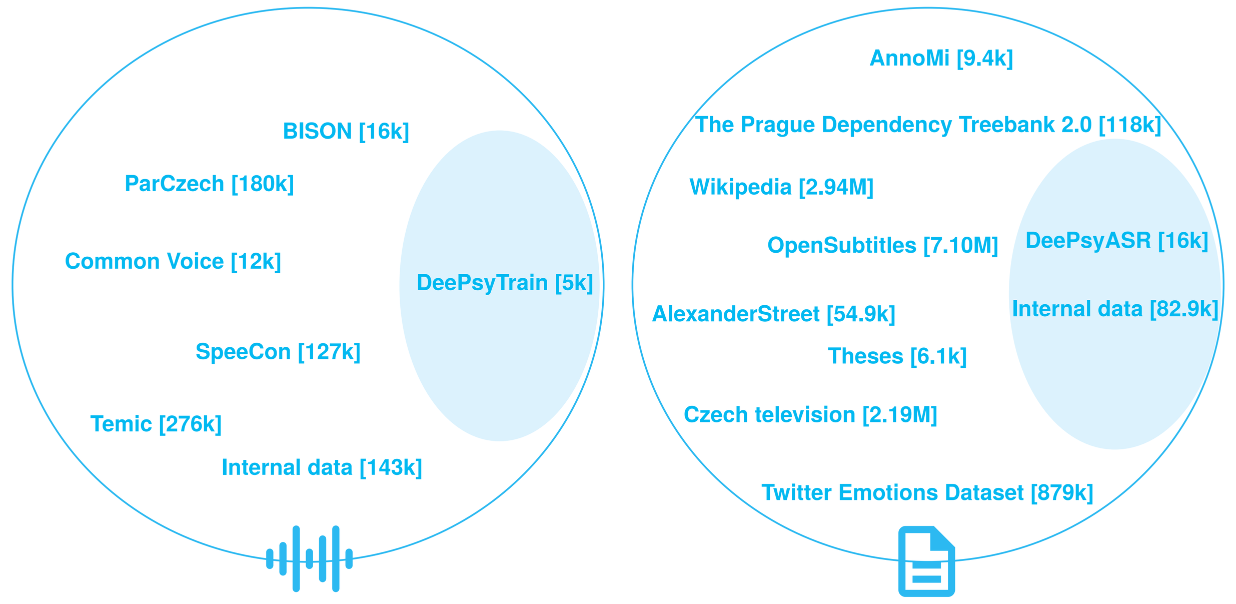

Figure 3. Overall training dataset consists of 921 labeled hours of speech and 13 million text sentences.

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 pretrain speech encoders **700 hours** of collected unannotated psychotherapy sessions, referred as **DeepSyUnsupervided**, were used. Downstream speech/text models were finetunned 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 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.

Model	Collar 0 ms			Collar 250 ms		
	DER [%] ↓	FA [%] ↓	M [%] ↓	DER [%] ↓	FA [%] ↓	M [%] ↓
vad_baseline [6]	10.63	9.35	1.29	5.87	4.67	1.20
Energy GMM	20.05	1.34	18.71	16.84	0.81	16.03
GPVAD [2]	8.92	**1.03**	7.90	6.05	**0.27**	5.78
PyanNet [1]	12.38	3.31	9.08	8.56	1.50	7.06
multilingual MarbleNet [3]	14.46	10.91	3.56	10.38	7.22	3.16
PyanNet	**6.61**	4.22	2.39	**2.99**	1.38	1.60
multilingual MarbleNet	12.74	12.42	**0.31**	8.54	8.25	**0.29**
CRDNN	9.50	7.42	2.07	5.58	4.00	1.58

Table 1. The error rate of pretained and finetunned (separated by horizontal line) 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 .

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 28.30% 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 Table2.

System	WER [%] ↓
CNN-TDNN-HMM	$28, 30$
XLS-R-300m	$31, 77$
+ 3 gram LM	$25, 12$
frozen XLS-R-300m + warm inited GPT 2	$29, 31$
XLS-R-300m + cold inited decoder	$27, 56$
+ beam decoding	$\mathbf{23, 47}$
Whisper-medium	$24, 25$

Table 2. Analysis of the error rates of the trained models on the DeePsyTest dataset.

Therapeutic Intervention Type Classification

To train and evaluate models for classifying types of therapeutic interventions, the DeePsy project created the **DeePsyInterventions** dataset. This dataset contains categories such as questioning, interpretation, reflection, confirmation, information, directives, self-disclosure, confrontation, and others, comprising over **14 thousand utterances**.

Category	F1 ↑
Questioning	0.77
Reflection + interpretation + information	0.81
Confirmation	0.76
Directives + confrontation	0.41

Table 3. Following the steps described in this thesis, a system based on the FERNET [5] model was finetuned to obtain results vidible in this table.

Conclussions

As part of this thesis, **significant improvements**, detailed in Table 4, in the tasks of speech activity detection and speech recognition and the steps that led to these improvements were presented.

In addition, the *end-to-end* **new systems** for diarization and overlapping speech detection **were trained**. **Current limitations** for trained models for sentiment classification and therapeutic interventions **were discussed**, and finally the **future steps have been presented**.

Task	System	Metric	Value	Rel. imp. [%]
Voice activity detection	PyanNet	FA+M [%] ↓	2.99	+49.06
Overlapping speech detection	PyanNet	F1 ↑	0.49	-
Diarization	Adapted VBx [4]	DER [%] ↓	6.10	-1.16
Sentiment classification	CZERT [7]	macro F1 ↑	0.45	-
Therapeutic intervention classification	FERNET	macro F1 ↑	0.47/0.69	-

Table 4. Summary table of the best results in the respective tasks.

References

[1] Hervé Bredin and Antoine Laurent.
End-to-end speaker segmentation for overlap-aware resegmentation, 2021.

[2] Heinrich Dinkel, Yefei Chen, Mengyue Wu, and Kai Yu.
Voice activity detection in the wild via weakly supervised sound event detection, 2020.

[3] Fei Jia, Somshubra Majumdar, and Boris Ginsburg.
Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection, 2021.

[4] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget.
Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks.
Comput. Speech Lang., 71(C), jan 2022.

[5] Jan Lehečka and Jan Švec.
Comparison of czech transformers on text classification tasks.
In Luis Espinosa-Anke, Carlos Martín-Vide, and Irena Spasić, editors, *Statistical Language and Speech Processing*, pages 27–37, Cham, 2021. Springer International Publishing.

[6] Oldřich Plchot, Pavel Matějka, Ondřej Novotný, Sandro Cumani, Alicia Díez Lozano, Josef Slavíček, Mireia Sánchez Diez, František Grézl, Ondřej Glembek, Mounika Veera Kamsali, Anna Silnova, Lukáš Burget, Francois Antoine Lucas Yang Ondel, Santosh Kesiraju, and A. Johan Rohdin.
Analysis of but-pt submission for nist lre 2017.
volume 2018, pages 47–53. International Speech Communication Association, 2018.

[7] Jakub Sido, Ondrej Prazák, Pavel Pribán, Jan Pasek, Michal Seják, and Miloslav Konopík.
Czert - czech bert-like model for language representation.
CoRR, abs/2103.13031, 2021.

This project is financed from the state budget by the Technology Agency of the Czech Republic and the Ministry of Industry and Trade within the Éta Programme.

ipoloka@fit.vut.cz