LEVERAGING PRETRAINED MODELS FOR AUTOMATIC SPEECH RECOGNITION IN PSYCHOTHERAPY SESSIONS

Bc. Alexander Polok
Ing. Pavel Matějka, Ph.D

T A
Č R

MUNI

DeePsy

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel @FIT 2023

MOTIVATION

Automatic analysis of therapeutic session

Feedback/supervision for therapists after the session

Detection of subtle nuances essential for in-depth analysis of a dialogue

Allow therapists to focus more on their patients and provide them with the highest level of care

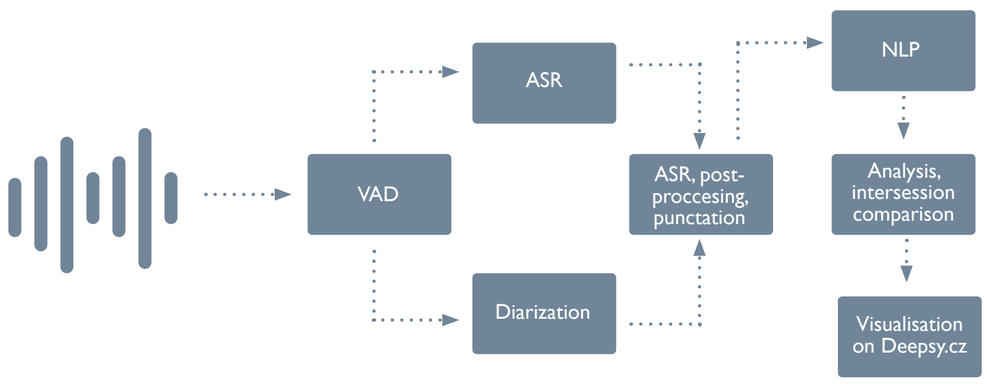

Figure 1: Design of the DeePsy system for automatic session analysis.

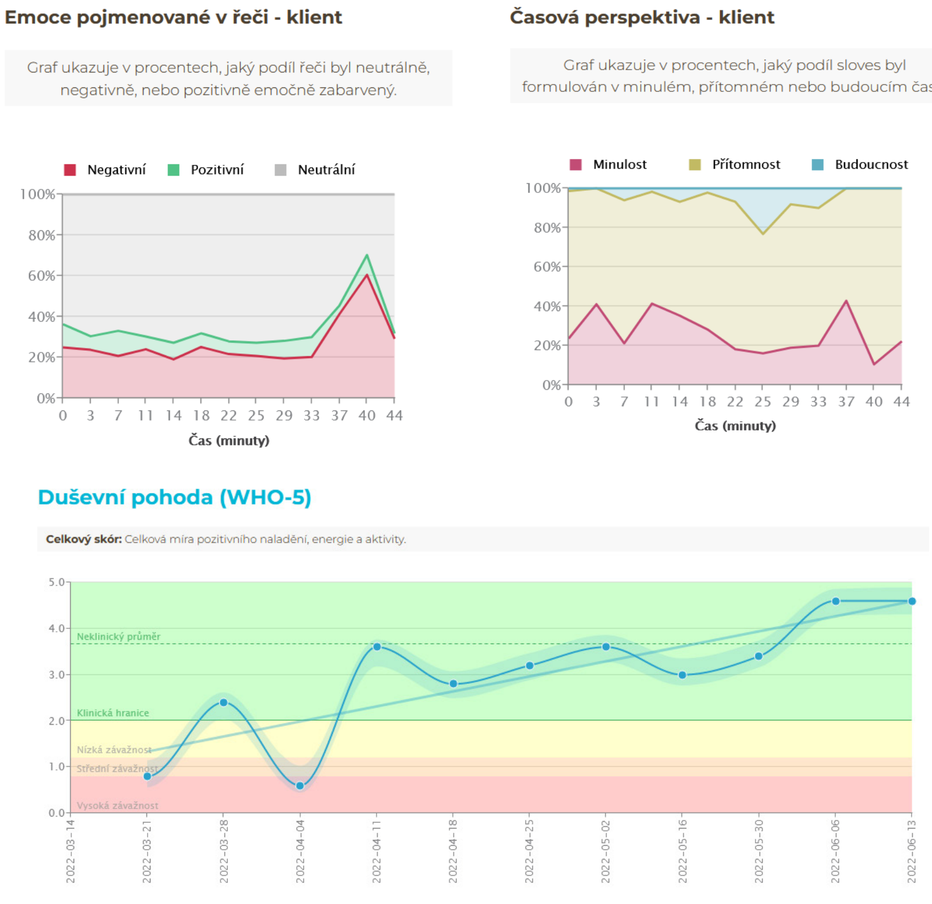

Figure 2: Examples of extracted features from the actual psychotherapeutic session within the DeePsy system.

WHAT IS AUTOMATIC SPEECH RECOGNITION (ASR)?

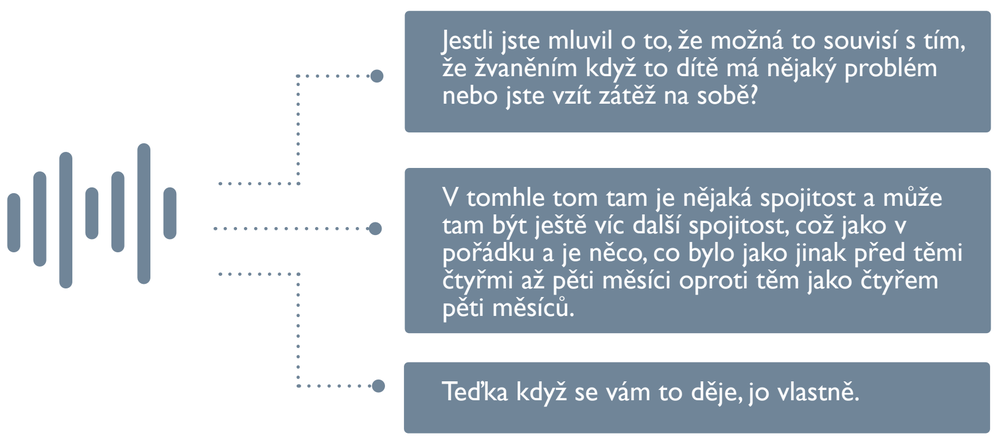

Figure 3: Instance of ASR system with a demonstrative input and textual transcriptions extracted from one of the sessions.

$$\text{WER} = \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Number of words in the reference}}$$

PROPOSED SOLUTION

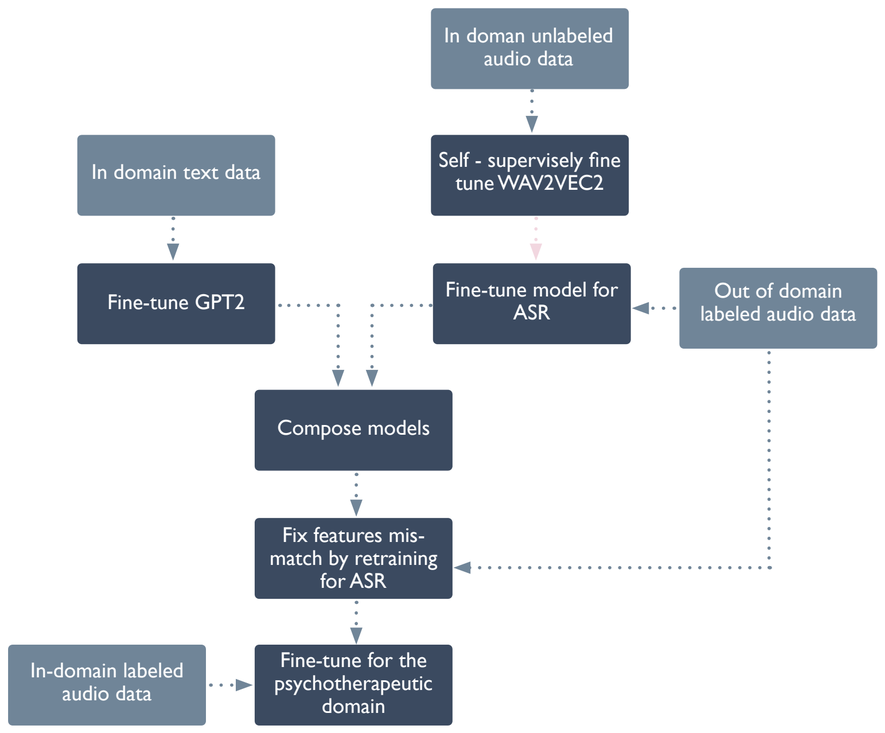

Figure 4: Diagram demonstrating proposed training protocol.

CONCLUSIONS

Adapted XLS-R on unlabeled DeePsy data and showed significant improvement of the recognition capabilities of correct speech unit among extractors by **8.40%** relatively.

Finetuned Wav2vec2 and Whisper models for the psychotherapeutic domain in the Czech language and designed a training protocol to enhance the ASR system's performance by **11.6%** WER relatively.

Trained and introduced speech and textual feature extractors that will be further incorporated into the emotion recognition, summarization, or classification of therapeutic intervention types.

Model	# parameters	WER [%]	CER [%]
CITRUS	95 mil	54,64	33,72
XLS-R	300 mil.	45,93	28,64
Whisper-base	74 mil.	52,40	32,03
Whisper-small	244 mil.	56,86	36,17

Table 1: Accuracy of fine-tuned models on the DeePsyTest dataset.

System	WER [%]
CNN-TDNN-HMM	28.3
XLS-R-300m	45,93
+ augmentations	40.76
+ 3 gram LM	32.03
+ 7 hours of in domain labeled data	25.12
+ GPT2 rescoring	25.01

Table 2: Gradual improvements of the system.

The development of the application is supported by the Technology Agency of the Czech Republic under the Eta Programme (grant no. TL03000049).

Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.