BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Leveraging Pretrained Models for Automatic Speech Recognition in Psychotherapy Sessions

Alexander Polok, Pavel Matějka*

Abstract
The DeePsy project aims to design and develop features that accurately model psychotherapeutic session dynamics, which can reveal subtle nuances essential for in-depth session analysis. However, these features are directly impacted by the accuracy of the preceding systems and are deemed unreliable with flawed Automatic Speech Recognition (ASR) systems. This work seeks to enhance the quality of psychotherapy session analysis by comparing several pre-trained ASR models applicable to the Czech language, adapting them to the psychotherapeutic domain, and developing a training protocol that effectively combines labeled/unlabeled out/in-domain textual and audio data to obtain the best ASR system possible. Audio and text feature extractors trained within this work are further utilized to develop a reliable assistive tool to help therapists professionally grow and provide better psychotherapy in the future. The enhanced ASR system and feature extractors can improve the accuracy and efficiency of psychotherapy sessions, allowing therapists to focus more on their patients and provide them with the highest level of care. The proposed training approach achieved an 11.6% relative improvement in Word Error Rate (WER) compared to the hybrid baseline system.

*xpolok03@stud.fit.vutbr.cz, matejkap@fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

The architecture of Automatic Speech Recognition (ASR) models has undergone a significant transformation with the emergence of deep neural networks. The originally widely used ASR systems using Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) [1] were gradually replaced by end-to-end (E2E) systems. The main advantage of this approach is a single objective function that is consistent with the system requirements for ASR as opposed to hybrid models, where the relevant components are optimized in isolation [2]. To explore the efficacy of E2E approaches, we conducted initial experiments on Czech audio data using the CITRUS [3] model, which is based on the WAV2VEC2 [4] architecture within a CTC [5] framework. Subsequently, we incorporated the cross-lingual XLS-R [6] and seq-2-seq Whisper [7] models to build more complex AED [8, 9] architectures. Results were compared against the hybrid CNN-TDNN-HMM [10] baseline.

2. Data and experimental setup

All of the presented experiments were conducted on the DeeePsyTest dataset, which was partially annotated as part of this work. The dataset comprises 11 online, five mobile recorded, and 32 psychotherapy sessions recorded on a ZOOM H2n dictaphone, varying from 4 to 7 minutes totaling 4.1 hours of audio containing 3.4 hours of speech. To conduct initial experiments, we collected the multidomain ASR corpus [11, 12, 13, 14], later referred to as ASRCorpora. This corpus contains 754 thousand training samples of varying lengths (2 to 20 seconds), totaling 921 hours. We enriched the dataset with 7.6 hours of annotated target domain sessions from the DeePsy project. All audio data was preprocessed by resampling to a frequency of 16,000 Hz and mixing to a single channel.

The textual data used in the experiments were collected from several online sources [15, 16, 17], further expanded with 1045k samples from in-house data, referred to as LMCorpora. The textual data was stripped of any characters not included in the expanded Czech vocabulary.

To train the audio models with adequate batch size, segments longer than 20 seconds and shorter than 0.1 seconds were excluded, resulting in a dataset of 754 thousand training samples, 80 thousand validation samples, 5 thousand domain samples intended for training, and 3 thousand test samples. The quality of the proposed ASR system is evaluated by the Word Error Rate (WER) [18] since it is the most standard metric for ASR.

3. Experiments

In order to classify or extract complex features from dialogues, it is crucial to have high-quality speech and text features. However, the best up-to-date hybrid system CNN-TDNN-HMM [10] supplemented with an n-gram language model [19] reached a high error rate of WER = 28.3%. To address this issue, we conducted experiments with models based on the Transformer [20] architecture.

Specifically, we fine-tuned WAV2VEC2 models with a (classification) linear layer of size 46 and models with the Whisper [7] architecture on the ASRCorpora. The results, summarized in Table 1, demonstrate the importance of the number of model parameters compared to the source domain of training data. Experiments also revealed that smaller variants of the Whisper model were unsuitable for this specific domain.

Although the ASRCorpora training set includes many samples, initial experiments indicated a domain mismatch between it and DeePsyTest. Therefore, we analyzed the effects of augmentations on speech recognition in psychotherapy. We applied standard regularization techniques such as frequency, and time masking, extended by time warping from the SpecAugment [21] library. Most importantly, since the therapy session recordings contain significant reverberation, Room Impulse Responses (RIRs) were synthetically created and added to the signals using the Pyroomacustics [22] library. Training with mentioned augmentations led to a relative improvement of 4.37% WER and 4.81% CER.

To further reduce errors, we introduced n-gram models to the decoding step of speech recognition since the WAV2VEC2 models do not have an explicit language model, and n-gram models are cheap in inference and training. We trained models varying from 2-4 granularity with the KenLM [23] library. This step lowered the WER to 32.03% with the 3-gram model. As expected, introducing the in-domain DeePsyTrain training dataset, containing 7 hours of annotated sessions, significantly reduced errors, with a WER of 25.12% and a CER of 12.46%. As data annotation was expensive and the introduction of an n-gram LM played a crucial role in error reduction, we fine-tuned GPT2 [24] to LMCorpora to overcome the modeling limitations of the 3-gram model. By combining acoustic, n-gram, and external attention-based models, we further reduced the WER to 25.01%. However, the improvement was minor since the GPT2 could not reliably assign scores to given hypotheses. Therefore, we removed the CTC layer of the XLS-R and incorporated GPT2 directly into the model as a decoder, leading to a 29.07% WER. We also conducted experiments on pretraining XLS-R on all DeePsy unsupervised data, which showed significant improvement in the modeling capabilities of the model by improving recognition capabilities of correct units among extractors by 8.40% relatively. This model will be the building block for the next training iteration. Table 2 summarizes the experiments conducted so far.

4. Conclusions

Our experiments showed that using pre-trained models, such as XLS-R, as a starting point for fine-tuning on in-domain data yields better results than training standard hybrid architecture from scratch. Furthermore, we experimented with various data augmentation techniques, analyzed the improvement gained by introducing in-domain data, and designed a protocol to enhance the system's performance to the best values. These findings highlight the importance of careful data curation and augmentation, as well as the benefits of leveraging pre-trained models to improve the performance of E2E ASR systems. Overall, our work contributes to developing a reliable assistive tool to enhance the quality of psychotherapy sessions and support therapists' professional growth.

Acknowledgements

The work was supported by the Technology Agency of the Czech Republic (TA CR) within Program Éta (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.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[2] Jinyu Li. Recent advances in end-to-end automatic speech recognition, 2022.

[3] Jan Lehečka, Jan Švec, Ales Prazak, and Josef Psutka. Exploring capabilities of monolingual audio transformers using large datasets in automatic speech recognition of czech. In *Interspeech 2022*. ISCA, sep 2022.

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[5] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. volume 2006, pages 369–376, 01 2006.

[6] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[10] Martin Kocour, Jahnavi Umesh, Martin Karafiat, Ján Švec, Fernando López, Jordi Luque, Karel Beneš, Mireia Diez, Igor Szoke, Karel Veselý, Lukáš Burget, and Jan Černocký. BCN2BRNO: ASR System Fusion for Albayzin 2022 Speech to Text Challenge. In *Proc. IberSPEECH 2022*, pages 276–280, 2022.

[11] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2020.

[12] Ondrej Bojar, Jonás Kratochvíl, and Peter Polak. Large corpus of czech parliament plenary hearings. In *International Conference on Language Resources and Evaluation*, 2020.

[13] Ondřej Glembek, Martin Karafiát, Lukáš Burget, and Jan Černocký. Czech speech recognizer for multiple environments. In *Radioeletronika 2006*, pages 1–4, 2006.

[14] Jan "Honza" Černocký, Jordi Luque, Carlos Segura, Martin Karafiát, Igor Szöke, Miroslav Skácel, Karel Veselý, Karel Beneš, Murali Karthick Baskar, Johan Rohdin, Pavel Matějka, Jan Švec, Michal Veselý, and Pavel Krajča. Big speech data analytics for contact centers BISON.

[15] Rudolf Rosa. Plaintext wikipedia dump 2018, 2018. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[16] Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague dependency treebank 3.0, 2013.

[17] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[18] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Marieum Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, and Paavo Alku. *Introduction to Speech Processing*. 2 edition, 2022.

[19] Dan Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[21] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data aug-

mentation method for automatic speech recognition. In *Interspeech 2019*. ISCA, sep 2019.

[22] Robin Scheibler, Eric Bezzam, and Ivan Dokmanic. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018.

[23] Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.