BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

System for Automatic Recognition of Czech Dialects

Bc. Ondřej Odehnal

Abstract

This work is about dialect identification (DID) and langauge identification (LID) using large pre-trained self-supervised (SSL) models such as WavLM.

We adapt a WavLM Base+ model via multi-headed factorized attentive pooling (MHFA) and fine-tune it sequentially on both VoxLingua107 and the NAKI Czech dialect dataset.

Our best model achieves 80.3 % accuracy on the four main dialect groups and 77.1 % on the 13 dialect subgroups, outperforming previous ECAPA-TDNN baseline. Note that the performance of the model is dependent on the dataset and NAKI still requires more data cleaning.

This project is part of the NAKI program and will support future research to preserve, archive, and protect Czech regional dialects by demonstrating effective neural approaches for fine-grained dialect recognition.

*xodehn09@vut.fit.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Motivation: Czech dialects form an integral part of national and cultural identity, yet automatic tools for their recognition remain underdeveloped. Accurate dialect identification aids in preserving linguistic heritage and could enable dialect-aware speech technologies.

Problem definition: We address automatic recognition of 13 Czech dialect subgroups in 4 main regional groups, using only raw audio from long, variable quality interviews between single or several respondents (target group) and interviewers (considered noise).

Existing solutions: Prior work in LID leverages convolutional networks (ResNet) or TDNN back-ends (ECAPA-TDNN [1]) trained from scratch on large corpora. Self-supervised models (e.g. WavLM [2]) show promise in LID but have not been applied to fine-grained dialect tasks.

Our solution: We employ a self-supervised WavLM Base+ [2] frontend, attach a multi-head factorized alert (MHFA) pooling backend [3], and first fine-tune the model as LID classifier of the 107-language *VoxLingua107* [4] dataset, then do domain adaptation and train the model as dialect classifier (discarding the last projection layer) on the NAKI Czech dialect dataset (the data set is not open, but has an extensive overview in the following methodology [5]).

Contributions

- Achieving a LID accuracy of 94.97 % on VoxLingua107 dev dataset, which is better than the freely available models in HuggingFace [1].
- Achieving overall 77.12 % acuraccy for dialect subgroups identification on the NAKI dataset.
- Pre-processing of the "raw" NAKI dataset (data exploration, cleaning, dataset splitting, etc.).

2. Language & Dialect Identification

The task of language identification (LID) does not need an introduction; however, dialect identification (DID) is a captivating challenge that has not yet been fully explored. DID systems aim to preserve the local linguistic heritage and enable truly personalized speech technologies. Just consider how the communication speech patterns of a Moravian speaker differ from those of a Bohemian, or more generally, how any dialect speaker in your chosen language brings unique nuances and requirements.

The focus of this work is on the Czech dialects that are divided into 4 main groups and 13 subgroups of dialects, as shown on the map of Czech dialects in Fig. 1 [6]. From 1960 there was an ongoing effort to collect dialectological data by the Institute for Czech Langauge (UJČ) which still continues. See the jamap.cz for more information.

The data were gathered through various environments and with varying recording devices. They consist of audio recordings between single interviewers or many interviewers and respondents. Typically, the session lasts several minutes (+20 on average) and the dataset has 1,711 recordings (number is still growing) of around 440 hours of speech data. Additionally, the interviewer is considered to be speaking with non-target dialect, and thus unfortunately polute the recordings.

This makes the dialect dataset, NAKI, both low resource and quite challenging to work with. However, the proximity of the DID to LID enables us to elevate large language datasets such as VoxLingua107 which has an order of magnitude more speech data. We use this data set to initially fine-tune the model for the LID task on VoxLingua107 and then perform domain adaptation on the NAKI dialect dataset.

3. Data cleansing

The NAKI data set is quite problematic for automated processing as described in Section 2. Thus, we detected and mutted overlapped segments, and performed diarization with Pyannote [7, 8]. We remove the least-speaking person with a simple heuristic that they should let the respondent speak more. The result of total and filtered speech is in Fig. 2 showcasing yet another problem of inbalanced classes.

4. Model Description

For our problem, we decided to use the large SSL pre-trained model, WavLM Base+. This model has already "heard" large amounts of speech and captures various levels of information per layer as shown in Fig. 3. Adapting only the last layers works well for automatic speech recognition (ASR); however, for tasks like speaker identification (SID), it is better to pull information from the lower layers [3]. We experimented with what is the best for LID in Section 6.

5. Model Training

The model training consists of fine-tuning the SSL model on the LID (20 epochs) task using the VoxLingua107 dataset, followed by domain adaptation for DID on NAKI data (20 epochs) as shown in Fig. 4. We also experimented with training solely on NAKI data. We trained the models with WeSpeaker [9], and used well-established techniques such as SGD with weight decay, learning-rate scheduling with warm-up, additive angular margin (AAM) softmax loss, data augmentation via MUSAN [10] noise and RIR [11] reverberation corpora, and layer-wise learning-rate decay. We trained both back-end and front-end models.

6. Experiments & Results

For LID fine-tuning, we achieved best performance with the WavLM model, MHFA back-end, and softmax (SM), Angular Softmax (AS) loss. The precision in VoxLingua dev is around 94,9 %, as shown in Table 1. This beats other freely available LID models [1].

The best models for the domain adaptation were for further training on NAKI data. The best overall performance was achieved with VoxLingua107 training, then NAKI, accompanied by MHFA backend + SM or AS. However, as shown in Fig. 5, the model with the SM loss function places more weight on layers very close to the embeddings after the CNN encoder. This is very close raw audio and hints at overfitting, thus we consider the model with AS loss to be more robust.

The summary of the accuracy for the different dialects in Fig. 6 demonstrates that the model has great accuracy in the dialects with more recordings. However, it performs poorly in certain dialects, such as Walachian with 35 % accuracy. This is either due to the lack of proper data, or they might be stronger similarity between the Walachian and Kopanice regions.

7. Conclusions

We demonstrate that large self-supervised models, when paired with attentive pooling and sequential fine-tuning, work well for fine-grained Czech dialect recognition. Our adapted WavLM achieves accuracy 77.12 % for dialect subgroups identification, a substantial improvement over conventional baselines. However, the dataset still requires further work to be done, namely the interviewers speech filtering, as described in Section 3. As part of the NAKI program, this work lays the groundwork for robust dialect-aware speech technologies and contributes to the preservation and study of Czech linguistic diversity.

References

[1] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin,

William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[2] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022.

[3] Junyi Peng, Oldrich Plchot, Themos Stafylakis, Ladislav Mosner, Lukas Burget, and Jan Cernocky. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification, 2022.

[4] Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition, 2020.

[5] Milena Šimečková, Bohumil Stupňánek, Martin Karafiát, Václav Voženílek, Andrea Vondráková, and Radim Nétek. Metodika pro převod strukturovaných znalostí z oboru dialektologie do strojového učení, 2025. Accessed: 2025-04-27.

[6] Jazyková paměť regionů. Mapa nářečí českého jazyka. https://www.jazykovapamet.cz/, 2024. Map of Czech dialects available for download. Accessed on 2025-04-27.

[7] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

[8] Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.

[9] Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[10] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *CoRR*, abs/1510.08484, 2015.

[11] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224, 2017.