Improved Dialect Recognition by Adaptation to a Single Speaker

Manuel Vogel[1], Guido Kniesel[1], Alberto Calatroni[1] and Andrew Paice[1]

[1]iHomeLab Think Tank and Research Centre for Building Intelligence, Lucerne University of Applied Sciences and Arts (HSLU), Technikumstrasse 21, CH-6048 Horw, Switzerland

Abstract

Voice assistants understanding dialects would help especially elderly people. Automatic Speech Recognition (ASR) performs poorly on dialects due to the lack of sizeable datasets. We propose three adaptation strategies which allow to improve an ASR model trained for German language to understand Swiss German spoken by a target speaker using as little as 1.5 hours of speaker data. Our best result was a word error rate (WER) of 0.27 for one individual.

1. Introduction

Automatic Speech Recognition (ASR) refers to the task of converting an audio signal into its written transcription and finds application, among others, in voice assistants. ASR performs well on so-called well-resourced[1] languages, while results on dialects, specifically Swiss German, are poorer. This is particularly inconvenient for the acceptance of applications involving smart assistants for elderly people, for whom it might be a big nuisance to switch to Standard German. ASR for Swiss German is challenging for several reasons:

1. Swiss German has no standardized written form and Standard German is the output of choice, meaning that the system must provide speech translation (ST) rather than mere recognition. For example, the German expression «wollen wir» could be pronounced and written in several different variants in Swiss German, e.g. «wömmer», «wemmer», «wemmr» or «wämmer».
2. Swiss German dialects are diverse and not geographically well confined. Thus, creating regional models would be challenging.
3. The publicly available Swiss German datasets are few and small compared to the corpora for other languages. Training on thousands of hours of data to account for variability is not possible.

The contribution of this work is an exploration on how person-specific data can be used to tailor known models towards better performance for a specific individual instead of trying to learn several different dialects with a single model. We investigated different approaches of training/fine-tuning and we assessed the performance of a pre-trained model adapted on a single speaker.

2. Related Work

In recent studies, the application of end-to-end ASR models (from raw audio to the words) based on deep neural networks has shown a considerable performance boost. To achieve good results, a considerable amount of training data is needed [1]. In the case of Swiss German, there is a lack of enough data, variability and the appropriate ground truth. An exception is the recently published Swiss Parliaments Corpus (SPC), which we use in our work [2].

The two noteworthiest end-to-end architectures are `Conformer` [3] and `wav2vec2` [4]. The latter achieved the best WER to date in German ASR (WER 0.057). Therefore, we chose `wav2vec2` as starting point.

ASR systems for low-resource dialects have lately attracted some attention [5, 6, 7, 8, 9]. When looking at Swiss German, we find the work of Plüss et al. [2], who claim a WER of 0.289 using a `Conformer` model on a the SPC dataset. Other researchers combined the SPC with a proprietary dataset to train various ST systems and achieved a WER of 0.5 when using only the SPC dataset [10]. A further approach achieved a WER of 0.39 on the SPC by training a model on a German dataset, transfer learning to the SPC enhanced with a proprietary internal dataset and refining the classification with a re-scoring [11]. Our evaluation yields results similar to Plüss et al [2], even if a direct comparison is not possible, and gives interesting insights about different fine-tuning strategies.

3. Materials and Methods

We here describe the baseline model, datasets and adaptation approaches.

SwissText 2022: Swiss Text Analytics Conference, June 08–10, 2022, Lugano, Switzerland

manuel.vogel@hslu.ch (M. Vogel);
guido.kniesel@hslu.ch (G. Kniesel);
alberto.calatroni@hslu.ch (A. Calatroni);
andrew.paice@hslu.ch (A. Paice)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]Well-resourced refers to the availability of abundant labeled data corpora to train machine learning algorithms.

3.1. Model

We base our work on a pre-trained `wav2vec2` model[2] available from the HuggingFace[3] AI community. We denote this model as `baseline`. The model topology consists of convolutional layers which map the raw audio to latent quantized speech representations and a Transformer structure which maps to context representations. The first pre-training stage involves self-supervised learning and therefore does not need labeled data [4]. For the `baseline` model, pre-training is done with multilingual data (53 languages) to learn language-independent speech units, followed by supervised training with German data, since Swiss German has strong similarities with German.

3.2. Dataset

In our experiments we used the Swiss Parliaments Corpus (SPC), a Swiss German dataset that contains recordings and transcriptions of the cantonal parliament of Bern (Grosser Rat Kanton Bern) [2]. It contains 293 hours of audio by 198 speakers and represents the biggest Swiss German speech recognition dataset to date. We use a subset of the SPC containing only samples with a high alignment between text and audio[4]. The audio files contain mostly Swiss German speech, whereas the labels (transcriptions) are in Standard German. We chose SPC mainly because of its size. In comparison to existing Swiss German datasets, such as ArchiMob [12], SPC has more audio data, which allows us to experiment better with various sizes of the single-speaker datasets. In addition, we recorded a new small dataset from a speaker unrelated to the SPC dataset, which allows us to test our approaches on another context. It is based on utterances of the Voxforge speech corpora[5]. We call this the external speaker (shortened: «ext»).

3.2.1. Dataset Partitioning

From the SPC dataset we created convenient partitions for the experiments. In the original corpus we identified the five speakers that have the biggest amount of data. These are the speakers with IDs 82, 145, 177, 186 and 207. Together with our own small external dataset, this yields six datasets. We refer to them as Single Speaker Corpora (SSC). For the approaches which involve a training step with multiple speakers, we extracted a subset of the SPC which excludes the speakers identified above (`SPC-without-top5`). Among the single speakers, the one with the least amount of data has around 1.5 hours of audio; therefore, for a fair comparison, we limited all six SSCs to 1.5 hours.

3.3. Approaches to Single Speaker Adaptation

Our goal is to adapt the baseline model to perform ASR satisfactorily with a single target speaker. We propose three approaches, which we evaluated on six SSC:

1. Supervised training of the `baseline` model on the SPC excluding all target SSCs.
2. Fine-tune the `baseline` model with data from the target SSC.
3. Combine the two previous approaches by training the `baseline` model on SPC excluding all target SSCs, then fine-tune with the target SSC.

In addition, we also evaluate the `baseline` model on all target SSCs. The three approaches are visualized in Figure 1.

4. Results

We conducted several experiments in line with the approaches described in Section 3.3 and evaluated the models against held-out test data, reporting the word error rate (WER). We show the results in Table 1. The best approach for adapting to a single speaker is the last one, i.e., to first train the model on Swiss German data from several speakers and then on the corresponding target speaker dataset. Interestingly, fine-tuning the baseline model only with the target single speaker datasets gives worse results compared to training a model on multiple speakers (`SPC-without-top5`). However, it is important to note that the SSCs contain only 1.5 hours of data, whereas `SPC-without-top5` contains around 176 hours.

The individual improvement of the adaptation depends on the speaker and varies between 1% and 4% on the five SPC speakers and reaches a notable 14% on the external speaker. Speaker 82 has the highest WER when evaluated with the base model but the lowest WER when fine-tuning with multiple Swiss German speakers and/or the single speaker dataset of speaker 82. In contrast, speaker 207 has the lowest WER when evaluated with the base model, but the highest WER using the other three approaches. The reasons for this behaviour could not be fully determined and further investigations are future work.

4.1. Influence of Training Data Amount

Increasing the data for the single speaker training has not led to a significant reduction of the WER. When training the model resulting from the second approach

[2]The exact model used is `wav2vec2-large-xlsr-53-german`.
[3]https://huggingface.co/
[4]Intersection over Union (IoU) > 0.9, `train_0.9`
[5]http://www.voxforge.org/de/downloads

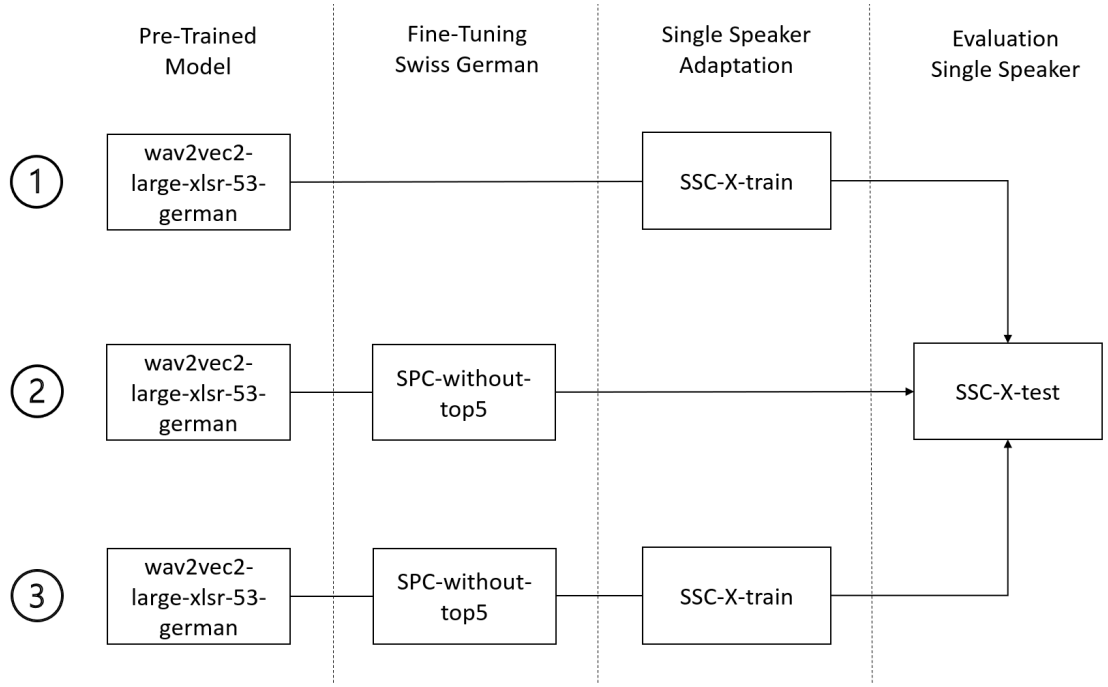

Figure 1: Visualization of the three adaptation approaches. `SPC-without-top5` denotes the SPC without all five target speakers and X denotes a speaker ID.

Approach	Fine-Tuning	82	186	207	177	145	ext
Baseline	-	0.90	0.89	0.82	0.89	0.86	0.82
1	SSC	0.56	0.67	0.60	0.64	0.60	0.50
2	SPC	0.31	0.42	0.44	0.35	0.38	0.44
3	SPC+SSC	**0.27**	**0.39**	**0.41**	**0.34**	**0.36**	**0.30**

Table 1
Word error rates (WER) for different speakers and training strategies. The numeric column headers are the SPC speaker IDs and ext denotes the external speaker. «SSC» and «SPC» stand for Single Speaker Corpus and Swiss Parliaments Corpus respectively. The first row shows the results of the pre-trained model.

with six hours of audio from speaker 82 instead of only 1.5 hours the WER decreases only by 2%. An identical improvement is observed when training with four hours of speaker 207 instead of 1.5 hours. Decreasing the time used for single speaker training does increase the WER: When training with a third of the external single speaker dataset, the WER increases by 4% and when training with a sixth, the WER increases by 5%.

4.2. Impact of Multi-Speaker Fine-Tuning

A remarkable result is the impact of Swiss German fine-tuning before the single speaker adaptation. Training the `baseline` model on the full SSC of speaker 82 (six hours), it achieves a WER of only 0.44. In comparison, training the model first with `SPC-without-top5` and then fine-tuning with 1.5 hours of speaker 82, achieves a WER of 0.27 on the same test set, giving an improvement of 17%. Training the `baseline` model on `SPC-without-top5` and then on 0.25 hours of data of the external speaker still performs better than using the model trained only on `SPC-without-top5` and the model trained only with 1.5 hours of data of the external speaker.

4.3. Limitations

One limitation is the prevalence of one specific dialect (Bernese) in the SPC. Furthermore, the SPC was recorded in a parliament and has therefore a certain bias in terms of content. The results can also be influenced by the combination of the chosen metric and the ground truth. For instance, if the audio contains the phrase *session vom september* and the label is *septembersession*, the WER increases if the model predicts the former phrase, even if

the two options are semantically identical. In addition, Swiss German does not have a past simple tense. Consequently, if the label is written in past simple, there is a significant difference in the structure of the spoken sentence and the ground truth.

5. Conclusion

We presented three possible strategies to adapt a pretrained ASR model based on `wav2vec2` to enhance the recognition on a single Swiss-German-speaking individual. The best strategy appears to be training a baseline model with multiple Swiss German speakers and in a second phase fine-tuning with a small amount of data from the target speaker. With this strategy, the WER for six speakers ranges between 0.27 and 0.41. The improvements of each approach on an external speaker and five SPC speakers are similar.

6. Outlook

Our adaptation approaches were tested only on one kind of model. It would be interesting to extend the evaluation to different models and examine if the behaviour and results are similar. Furthermore, the evaluation sample size n=6 is not quite representative considering the diversity of Swiss German. An evaluation containing more speakers will allow more solid claims.

References

[1] D. Jurafsky, J. H. Martin, Speech and Language Processing (3rd Edition Draft), USA, 2021. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf.

[2] M. Plüss, L. Neukom, C. Scheller, M. Vogel, Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus (2021). URL: http://arxiv.org/abs/2010.02810. arXiv:2010.02810.

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented transformer for speech recognition, 2020. URL: https://arxiv.org/abs/2005.08100. doi:10.48550/ARXIV.2005.08100.

[4] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations (2020). URL: http://arxiv.org/abs/2006.11477. arXiv:2006.11477.

[5] J. Sun, G. Zhou, H. Yang, M. Wang, End-to-end tibetan ando dialect speech recognition based on hybrid ctc/attention architecture, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 628–632. doi:10.1109/APSIPAASC47483.2019.9023130.

[6] R. Imaizumi, R. Masumura, S. Shiota, H. Kiya, Dialect-aware modeling for end-to-end japanese dialect speech recognition, in: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 297–301.

[7] Y. Zhao, J. Yue, X. Xu, L. Wu, X. Li, End-to-end-based tibetan multitask speech recognition, IEEE Access 7 (2019) 162519–162529. doi:10.1109/ACCESS.2019.2952406.

[8] Y. Zhang, M. Ablimit, A. Hamdulla, Error correction based on transformer lm in uyghur speech recognition, in: 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), 2021, pp. 204–207. doi:10.1109/PRML52754.2021.9520740.

[9] L. Pan, S. Li, L. Wang, J. Dang, Effective training end-to-end asr systems for low-resource lhasa dialect of tibetan language, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 1152–1156. doi:10.1109/APSIPAASC47483.2019.9023100.

[10] A. Khosravani, P. N. Garner, A. Lazaridis, Learning to translate low-resourced swiss german dialectal speech into standard german text, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 817–823. doi:10.1109/ASRU51503.2021.9688249.

[11] Y. Arabskyy, A. Agarwal, S. Dey, O. Koller, Dialectal speech recognition and translation of swiss german speech to standard german text: Microsoft's submission to swisstext 2021, 2021. URL: https://arxiv.org/abs/2106.08126. doi:10.48550/ARXIV.2106.08126.

[12] T. Samardžić, Y. Scherrer, E. Glaser, ArchiMob - a corpus of spoken swiss german, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), 2016, pp. 4061–4066. URL: https://aclanthology.org/L16-1641.