Domain Adversarial Training for German Accented Speech Recognition

Carlos Franzreb[1], Tim Polzehl[2]
[1] DFKI, 10559 Berlin, Deutschland, Email: carlos.franzreb@dfki.de
[2] DFKI, 10559 Berlin, Deutschland, Email: tim.polzehl@dfki.de

Introduction

The remarkable performance of current automatic speech recognition (ASR) models does not always translate to other accents, and is even worse for accents that are not present in the training data. Phonetic differences among accents still pose challenges to ASR models [10]. Speakers also vary regarding their grammar and their vocabulary, which complicates the task of the language model that either accompany or are embedded in ASR models [2]. Overcoming this issue is important for everyone to profit from their use. Accented Speech Recognition (AccSR) addresses the challenge posed by accents for ASR models, with the goal of improving their robustness to accents instead of aiming to improve the overall performance of the model.

[14] analyses how accent information is represented by an ASR model and proposes the use of auxiliary losses to help the model cope with accents. This was first attempted in a multi-task learning (MTL) fashion [11]. MTL consists of adding a secondary task (or more), whose optimization benefits the primary task. It has been successful in many scenarios, presumably due to the guidance provided by the secondary task, which directs the shared layers towards meaningful features [15]. The secondary task can be added to the ASR model adversarially, meaning that the ASR model is encouraged to discard accent information. Previous work ([16], [5]) with this setting assumed that there were no transcriptions for the accented speech. Given the increasing availability of accented speech, our goal here is to assess whether adversarial training is useful when there are transcriptions for the accented speech. We use the German Common Voice dataset [3], splitting the accented data into seen and unseen accents. Seen accents are used to train the model, both with their transcripts and their accent labels. Unseen accents are used to evaluate the ASR model's generalization ability on accents that it has not seen during training, as this is one of our main concerns.

We first perform optimization experiments and find that using separate learning rates for the two models is the most effective way of leveraging the feedback of the accent classifier. We then train the model adversarially with an accent classifier, with two kinds of reversal layers. One of them reverses all gradients, as is usually done in the literature, and the other one reverses only the gradients of accented speech. The intuition behind this new reversal layer is that we are interested in changing the representation of accented speech to improve its performance, but standard speech already performs well, and its representation should remain intact. Our results show that this new reversal layer improves the performance of seen accents. We compare the adversarial training with normal training, where all data is used to train an ASR model without a secondary task. Normal training already introduces invariance to accents, confirming the results of previous work on a related task [1].

Related work

Our main reference is a paper that applied domain adversarial training (DAT) to AccSR [16]. The authors achieved a 7% character error rate reduction without using the transcriptions of the accented speech. They also trained a model with the transcriptions of the accented speech; the results are similar to those achieved with normal training, when no accent classifier is used. The main difference between their scenario and ours is the accuracy of the ASR models, as they train a small TDNN from scratch, whereas we use a large Conformer Transducer which was already trained extensively. Another study [5] also applied DAT to AccSR, but with the English Common Voice dataset and a pre-trained ASR model (Nvidia's English QuartzNet). DAT slightly improved the performance of the ASR model, but they did not perform experiments with the transcripts of the accented data.

Experiments with the number of shared layers and the task weights have been performed by two studies on MTL for AccSR. [21] shows that the ASR model performs best when the weight assigned to the accent classifier is smaller, although the classifier loss was one order of magnitude smaller than the ASR loss. The same results were reported by [19]. [1] uses speaker labels and transcripts to compare MTL and DAT combined with speaker classification, using normal training without a secondary task as a baseline. Contrary to [21]'s results, the speaker classification accuracy degrades with the number of shared layers, suggesting that the ASR model already develops speaker invariant representations. Their results also show that neither MTL nor DAT offer any benefits over normal training. We perform similar experiments, but with a larger and pre-trained ASR model, and regarding accents instead of speakers, which differ more significantly from one another.

Setup

Domain Adversarial Training (DAT) increases a model's robustness to different domains by training it adversarially with a domain classifier [6]. We want our ASR model to be robust to accented speech, and therefore train it together with an accent classifier as an adversary. The key component for performing DAT is the gradient reversal layer (GRL), which is prepended to the accent classifier.

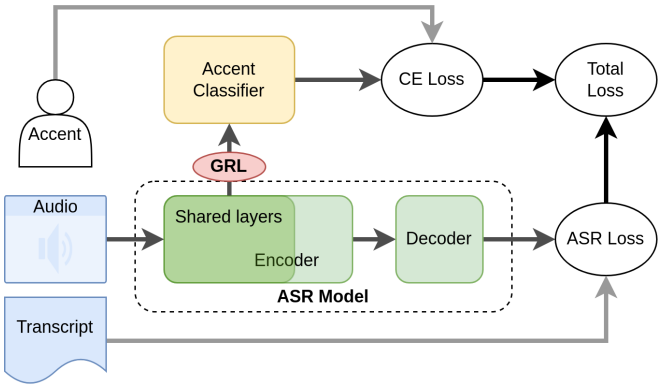

Figure 1: Diagram of the ensemble model

The GRL does nothing in the forward pass and negates the gradients in the backward pass. This negation inverts the classifier's goal; the ASR layers that precede the GRL are encouraged to remove accent information from their representations to increase the loss of the accent classifier. A diagram of this architecture is depicted in Figure 1. The classifier receives the output of an encoder block as input and predicts the accent of the representation. Its loss is then added to the ASR loss in a weighted sum, where the weights of the two losses are constrained to sum up to one. DAT was initially conceived for unlabeled data, which in our case means speech with accent labels and without transcripts. Here, we use both and train both models jointly.

We perform experiments on the Conformer Transducer trained by NVIDIA[1], which comprises 17 Conformer blocks in the encoder and has around 120M parameters in total. It has been initialized with its English counterpart and trained with over 2,000 hours of German data from Common Voice [3] and other datasets, achieving a WER of 4.93 on the test set of Common Voice 7. A study [20] shows that this model achieves the best WER on multiple datasets when compared with other Nvidia models, as well as Wav2vec 2.0 [4]. The Conformer Transducer architecture combines two concepts that have been very successful in speech recognition. The Conformer [9] introduces a convolutional layer into the Transformer block [18], combining their strengths to capture both the local and the global context of the input speech. Transducer models [8] improved the state-of-the-art by embedding a language model into the ASR pipeline, which conditions the model's predictions on the meaning of what is being transcribed.

Our work is based on NVIDIA's Nemo toolkit [12]. The accent classifier is appended to the desired layer of the ASR encoder. We use the same accent classifier for all our experiments: a binary classifier comprising two fully connected layers with ReLU activation of sizes 1024 and 256, followed by two output neurons. The classifier loss is the cross entropy loss. We experimented with adding dropout to the classifier, but it did not improve performance. Training data is augmented with SpecAugment [13]. The code, weights and results of our experiments are available online[2].

We use the German Common Voice dataset [3] for our experiments (version 10). We have split the data into accents and kept those for which the most data was available. Over 90% of our dataset (595 hours) belongs to the standard accent. We use 350 hours of it to fine-tune the model before performing the experiments. The rest, as well as Swiss and Austrian data, is split into training and test data. Together with German, they are the *seen accents*. The rest of the accents, shown in Table 1, comprise at most 2 hours of data, which is not enough to split into two subsets. We therefore use them only for evaluation, to assess the generalization ability of our trained models. We call them *unseen accents*. The WERs of NVIDIA's Conformer Transducer are shown in Table 1; the worst performing accent's WER (Lower Rhine dialect) is almost three times as large as the WER of the best accent (Alemannic). Our goal is to reduce this gap.

Table 1: Hours of data and WER of Nvidia's Conformer Transducer for each accent

Accent	WER	Test	Train	Pretrain
German (Std.)	3.4	49.1	196.1	350
Austrian	3.2	6.4	25.6	
Swiss	4.2	4.5	17.9	
French	4.6	2.1		
Italian	2.3	1.6		
Russian	4.6	1.1		
Alemannic	2.1	1.1		
Lower Rhine	5.7	1.1		
American	5.6	0.5		
British	5.1	0.2		
Canadian	5.0	0.2		

One-way reversal layer

During backpropagation, the GRL negates all gradients from the accent classifier before passing them to the ASR encoder. We experiment with a new reversal layer where only the gradients from accented speech samples are negated, as we want to improve the ASR model's representation of accented speech, while keeping the representation of standard speech intact. Standard DAT aims to swap the domains: by negating all gradients, the ASR model is encouraged to change the representations of both standard and accented speech. The accent classifier demands standard speech to be represented as accented speech, as well as the other way around. One-way DAT encourages the ASR model to represent accented speech as standard speech, but also to keep the representation of standard speech.

Experiments

Before performing the experiments, we fine-tune the ASR model with the pre-training dataset for 20 epochs to discard domain adaptation as one of the causes of improvement in later experiments, as the ASR model was

[1] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_de_conformer_transducer_large`

[2] `https://github.com/carlosfranzreb/accsr-daga23`

trained with several datasets, each with different recording environments, sample rates, etc. We then pre-train the accent classifier for 20 epochs while keeping the ASR model frozen, to ensure that its feedback to the ASR model is meaningful from the start [5]. We have experimented with four different branching points, concluding that branching on the eleventh encoder block works best, although the differences among branching points are small. These results agree with previous work ([19], [21]).

Hyperparameter optimization

We first perform optimization experiments, looking for the best way of leveraging the feedback of the accent classifier. Previous work ([21], [19]) experiments with the task weights, concluding that assigning a small weight to the classifier works best. We experiment with three different pairs of weights and achieve the same results: increasing the classifier weight increases the WER considerably. Using a different learning rate (LR) for the classifier, while keeping the ASR model's LR constant at the authors' recommended value ($1e-5$), is more effective. Assigning a larger LR to the classifier decreases the overall WER, with a relative improvement of 13% for the seen accents and 15% for the unseen accents.

Adversary configurations

Table 2 shows the WERs achieved with each technique after 40 epochs. No AC means the ASR model was trained without an accent classifier. This approach is the best for the standard accent, one other seen accent, and half of the unseen accents, showing that normal training already improves the model's robustness to accents. Previous work reached the same conclusion when investigating different speakers instead of accents [1]. However, both DAT approaches offer better performance on unseen accents on average. The difference between normal and adversarial training is largest for the worst-performing accent (American), where DAT is better than normal training by 0.22. The one-way reversal layer is better than reversing all gradients for seen accents; leaving the standard representation unchanged allows the model to optimize the representation of the other accents.

We use the MAPSSWE test [7] to measure the statistical significance of adversarial training. The test scores are above 70% for all but three accents, meaning that normal and adversarial training usually make the same prediction errors. This high correlation between the models is partly caused by the high label noise of Common Voice, which is a crowd sourced dataset without any filtering. Also, the WER of all models is already very low; they barely make any errors, which again brings their behavior closer. We therefore need to repeat this experiments with a larger dataset, where label noise and overfitting play lesser roles.

Distances between encoder representations

Finally, we investigate how each training method changes the encoder representations of the different accents. We do so by computing the Earth's Mover's distances (EMDs) between the representations of each accent and the standard representation. It is more expressive than computing distances between means, as it considers the whole distributions; previous work on accented speech has also used it to compare accent representations [14]. Given the large size of the encoder representations (512x512), we first reduce their dimensionality with t-SNE [17] to two dimensions.

Table 2: WERs achieved by the different training techniques

Accent	No AC	DAT	One-way
German	**1.60**	1.65	1.65
Austrian	**1.53**	1.63	1.61
Swiss	2.67	2.70	**2.63**
British	**3.71**	4.19	**3.71**
Italian	**1.24**	1.31	1.35
Alemannic	**1.13**	1.19	1.20
French	**1.92**	2.00	2.04
Lower Rhine	4.20	**4.11**	4.17
American	4.60	**4.38**	4.47
Canadian	3.27	**2.76**	2.93
Russian	2.29	2.13	**2.12**
Mean	2.56	2.55	**2.54**
Seen mean	**2.10**	2.16	2.12
Unseen mean	2.79	2.76	**2.75**
Worst	4.60	**4.38**	4.47

Table 3: EMD distances from the accents to the standard accents at the branching point.

Accent	Baseline	No AC	DAT	One-way
Seen	3.9	2.0	1.2	1.9
Unseen	7.2	3.8	3.9	3.5

The EMDs for each experiment are similar, as shown in Table 3. Normal training brings the accent representations closer to the standard representation as much as DAT. However, these distances are not indicative of WER. For example, DAT is the worst for seen accents, but its representations are closest to the standard accent. Still, all experiments improve upon the baseline, and unseen accents are further apart from the standard accents than seen accents, which is expected. Using accent classification, although it is more expensive to compute, may be a more accurate indicator of WER. We will test this hypothesis in future work.

Conclusion

Our goal was to evaluate the usefulness of adversarial training for accented speech recognition. Normal training without an adversary already develops accent-invariant representations and improves the ASR model's performance on unseen accents. Adversarial training only yields slight improvements, but this maybe because of the small size of the dataset and the lack of accented speech in the training set. We plan to perform the same experiments with English Common Voice, which is considerably larger.

Acknowledgements
This research has been funded by the Federal Ministry of Education and Research (BMBF), Germany, Project PYSA 16SV8850

References
[1] Y. Adi, N. Zeghidour, R. Collobert, N. Usunier, V. Liptchinsky, and G. Synnaeve. To Reverse the Gradient or Not: an Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3742–3746, May 2019. ISSN: 2379-190X.

[2] A. Aksënova, Z. Chen, C.-C. Chiu, D. van Esch, P. Golik, W. Han, L. King, B. Ramabhadran, A. Rosenberg, and S. Schwartz. Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data. *arXiv preprint arXiv:2205.08014*, 2022.

[3] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[5] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau. Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning, Mar. 2021. arXiv:2103.05834 [eess].

[6] Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation, Feb. 2015. arXiv:1409.7495 [cs, stat].

[7] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 532–535 vol.1, May 1989. ISSN: 1520-6149.

[8] A. Graves. Sequence Transduction with Recurrent Neural Networks, Nov. 2012. arXiv:1211.3711 [cs, stat].

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040. ISCA, Oct. 2020.

[10] A. Hinsvark, N. Delworth, M. Del Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin, N. Bhandari, and M. Jette. Accented Speech Recognition: A Survey, June 2021. arXiv:2104.10747 [cs, eess].

[11] A. Jain, M. Upreti, and P. Jyothi. Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning. In *Interspeech 2018*, pages 2454–2458. ISCA, Sept. 2018.

[12] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, P. Castonguay, M. Popova, J. Huang, and J. M. Cohen. NeMo: a toolkit for building AI applications using Neural Modules, Sept. 2019. arXiv:1909.09577 [cs, eess].

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019*, pages 2613–2617, 2019.

[14] A. Prasad and P. Jyothi. How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3739–3753, Online, July 2020. Association for Computational Linguistics.

[15] S. Ruder. An Overview of Multi-Task Learning in Deep Neural Networks, June 2017. arXiv:1706.05098 [cs, stat].

[16] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie. Domain Adversarial Training for Accented Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4854–4858, Calgary, AB, Apr. 2018. IEEE.

[17] L. van der Maaten and G. Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[19] T. Viglino, P. Motlicek, and M. Cernak. End-to-End Accented Speech Recognition. In *Interspeech 2019*, pages 2140–2144. ISCA, Sept. 2019.

[20] J. Wirth and R. Peinl. ASR in German: A Detailed Error Analysis. Apr. 2022.

[21] J. Zhang, Y. Peng, V. T. Pham, H. Xu, H. Huang, and E. S. Chng. E2E-Based Multi-Task Learning Approach to Joint Speech and Accent Recognition. In *Interspeech 2021*, pages 1519–1523. ISCA, Aug. 2021.