INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning

Nilaksh Das[1*], *Sravan Bodapati*[2], *Monica Sunkara*[2], *Sundararajan Srinivasan*[2], *Duen Horng Chau*[1]

[1]Georgia Institute of Technology, USA
[2]Amazon AWS AI, USA

nilakshdas@gatech.edu, sravanb@amazon.com

Abstract

Training deep neural networks for automatic speech recognition (ASR) requires large amounts of transcribed speech. This becomes a bottleneck for training robust models for *accented* speech which typically contains high variability in pronunciation and other semantics, since obtaining large amounts of annotated accented data is both tedious and costly. Often, we only have access to large amounts of *unannotated* speech from different accents. In this work, we leverage this unannotated data to provide semantic regularization to an ASR model that has been trained only on one accent, to improve its performance for multiple accents. We propose Accent Pre-Training (Acc-PT), a semi-supervised training strategy that combines transfer learning and adversarial training. Our approach improves the performance of a state-of-the-art ASR model by 33% on average over the baseline across multiple accents, training only on annotated samples from one standard accent, and as little as 105 minutes of *unannotated* speech from a target accent.

Index Terms: Accented speech recognition, domain adversarial training, transfer learning, semi-supervised learning

1. Introduction

Huge advancements in the field of automatic speech recognition (ASR) have been made in the recent years by leveraging deep learning techniques [1–6]. State-of-the-art ASR systems nowadays employ massive deep neural networks (DNNs), which are trained on tremendous amounts of annotated speech data. This revolution in the ASR community would not have been possible without undertaking arduous data collection and annotation efforts for human speech. However, in practice, there is usually limited diversity in the speech accent for large annotated datasets that are used to bootstrap ASR systems. Moreover, datasets with higher multiplicity of annotated accented speech are typically deficient in size, thus being rendered as inadequate for training large ASR systems from scratch. This degrades the generalization capacity of ASR systems with respect to accented speech examples. In this work, we aim to improve the accent robustness of a trained state-of-the-art ASR model in this low-resource setting for accent data.

While we have huge amounts of annotated data from certain locales (e.g., the largest LibriSpeech dataset has primarily US-accent), it is oftentimes the case in practice that we only have access to *unannotated* data from other locales as it is easier to collect. In this work, we focus on this setting to explore how much gain can be made without undertaking the expensive and tedious task of annotating accented speech. Several semi-supervised learning techniques have been proposed to tackle this scenario by combining annotated and unannotated examples to jointly learn robust inference models [7–9]. We propose to leverage domain adversarial training (DAT), one such semi-supervised approach [10] for our low-resource setting in the absence of annotated accented speech data. DAT is an adversarial training technique that enforces intermediate representations to be domain-invariant for different accented examples, and has been shown to improve the accent robustness of ASR models with limited annotated data [11]. In this work, we augment this approach by proposing a training strategy that incorporates transfer learning for further improving performance from DAT.

Transfer learning is a training paradigm that leverages a model trained for a particular *base* task, to reuse it as a starting point for training the model on another *new* task. This bootstrapping method is especially powerful when the new task has limited data, and is similar to the base task. The bootstrapped model would have some latent knowledge of the new task, and does not require large amounts of new data. In this paper, we leverage a state-of-the-art ASR model that has been pre-trained for speech recognition for a base task, and further train it with DAT for a new task in which we have no accented speech annotations, and have limited accented data overall. This mimics several practical scenarios in deploying real-world ASR systems. Our work makes the following major contributions:

- We improve upon the semi-supervised DAT technique, combining it with transfer learning. Specifically, we propose to bootstrap the discriminator to accurately classify accented speech from the intermediate feature space of a trained encoder, before applying the DAT training. We call this approach "Accent Pre-training" (Acc-PT).
- We evaluate our approach across multiple accents, and show that our proposed Acc-PT technique combined with DAT improves the performance of the state-of-the-art ASR model by 33% over the baseline across several accents on average, despite having been trained only with annotated speech samples for one standard accent (US), and as little as 105 minutes of *unannotated* speech from a target accent (Philippines).

2. Related Work

Domain adversarial training (DAT) is a popular semi-supervised approach used in several computer vision tasks [10, 12]. It has been shown to improve the robustness of ASR models in the low-resource setting, when limited annotated data is available [11, 13, 14]. DAT aims to make the intermediate latent space domain-invariant by adversarially training a discriminator and encoder, thus regularizing the downstream decoder weight parameters to be domain agnostic. Learning such a domain-invariant acoustic feature space has also been shown to improve

[*]Work conducted as an intern at Amazon AWS AI.

  http://dx.doi.org/10.21437/Interspeech.2021-1888

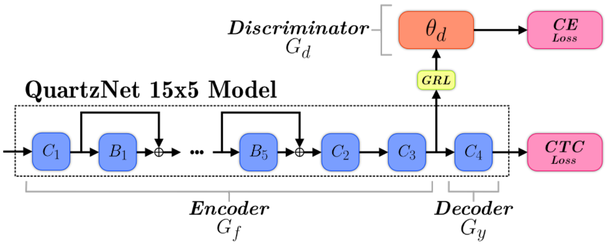

Figure 1: *DAT with QuartzNet 15x5 model.*

ASR robustness in noisy environments [15, 16] and for speaker verification tasks [17–19]. For accent robustness, this method is leveraged to train the end-to-end ASR model (encoder and decoder) with annotated data from a source accent, and simultaneously train the encoder and discriminator combination with unannotated accented data from target accents. While previous work trains the ASR model from scratch [11], we explore the added benefit of leveraging transfer learning to improve a pre-trained ASR model in the low-resource setting, where we have limited amount of annotated source accent data and unannotated target accent data. Furthermore, we propose the Accent Pre-training (Acc-PT) strategy that bootstraps the discriminator by pre-training on learned representations from the model, which is key to fully adapting DAT to work with transfer learning.

Several other approaches have been proposed recently to train accent robust DNNs for ASR. However, many of these methods may fail in the low-resource setting. A hierarchical grapheme-based technique is used to jointly learn grapheme and phoneme prediction tasks [20] using annotated data from multiple accents. A multi-task learning (MTL) approach trains the ASR model jointly with accented English and native language annotated speech from Spanish and Indian speakers [21]. Some MTL methods for improving accent robustness also take an inverse approach than DAT by increasing the domain variance of the intermediate latent representation, thus jointly learning an "accent aware" ASR model [22, 23]. However, to jointly train such MTL models, annotated samples are required from *multiple* accents, posing a stark contrast to the DAT approach which requires annotated samples from only the source accent. For example, MTL models [22, 23] train on annotated speech samples from 7 accents, whereas our DAT approach only trains on annotated speech samples from just one accent. This demonstrates the versatility of the DAT approach, especially considering the low-resource setting. Previous DAT work outlines this distinction with MTL, showing that DAT outperforms MTL with limited data resources [11]. Hence, in this paper, we specifically focus on further improving the DAT approach by combining it with transfer learning.

3. Approach

Our main idea is to leverage the annotated speech of a standard accent, which is abundantly available in general, to improve the robustness of a trained state-of-the-art ASR model on the unannotated speech of other unseen accents. In this work, we denote the standard accent data having seen annotations as $S = \{\boldsymbol{x}_i, \boldsymbol{y}_i, a_i = 0\}_{i=1}^{|S|}$ and the other accented speech samples with different accents having unseen annotations as $U = \{\boldsymbol{x}_i, a_i \geq 1\}_{i=1}^{|U|}$. Here, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the input speech spectrogram features and the respective transcriptions, and a_i is the corresponding accent for a given sample.

Table 1: *The QuartzNet 15x5 architecture.*

Block	K	c_{out}
C_1	33	256
B_1	33	256
B_2	39	256
B_3	51	512
B_4	63	512
B_5	75	512
C_2	87	512
C_3	1	1024
C_4	1	‖labels‖

3.1. End-to-end Speech Transcription with CTC loss

For our baseline, we use a trained QuartzNet 15x5 model [4], which is a state-of-the-art end-to-end neural acoustic model for ASR, and is made publicly available[1] by the authors. QuartzNet is based on the Jasper architecture [24], a fully convolutional (conv) model trained with Connectionist Temporal Classification (CTC) loss [25]. A representation of the model architecture can be seen in Figure 1.

The first layer of the QuartzNet 15x5 model is a 1D conv layer C_1 having a stride of 2, which is followed by a sequence of blocks. Each block B_i is repeated 3 times, and has residual connections between blocks. Further, a block B_i consists of the following group of 4 layers, the group being repeated 5 times within the block: 1) K-sized depthwise conv layer with c_{out} channels; 2) a pointwise conv operation; 3) a batch-wise normalization layer; and 4) a ReLU activation. The final stage of the QuartzNet model has 3 more conv blocks (C_2, C_3, C_4), where C_4 (having a dilation of 2) does the final decoding of labels. Table 1 gives more detailed architectural specifications of the blocks.

In this work, we use the QuartzNet model that has been trained on the 960-hour LibriSpeech train set [26], achieving a near state-of-the-art ASR performance with WER=3.90% on the "test-clean" split and WER=11.28% on the "test-other" split, without using any language model. The model was trained for 400 epochs using the NovoGrad optimizer [27].

Using a pre-trained QuartzNet model allows us to establish a transfer learning baseline, where ASR performance on LibriSpeech (a dataset with limited accent diversity) constitutes the base task. Our new task for transfer learning is to perform robust transcription on a dataset with multiple accents (e.g., Common Voice dataset, described in Section 4.1). Our main idea is to leverage the latent knowledge learned by the trained model for recognizing human speech and augment its generalization capacity by re-tuning it in a semi-supervised fashion in the absence of annotated accented data.

3.2. Domain Adversarial Training (DAT)

DAT is an adversarial training approach that aims to learn an intermediate latent feature space that is domain-invariant. A DAT model consists of 3 main components: 1) the feature encoder network G_f with parameters θ_f, such that $\boldsymbol{f} = G_f(\boldsymbol{x}; \theta_f)$, where $\boldsymbol{x}$ are the input speech features; 2) the domain classifier (discriminator) network G_d with parameters θ_d, such that

[1] https://ngc.nvidia.com/catalog/models/nvidia:quartznet_15x5_ls_sp

$a = G_d(\boldsymbol{f}; \theta_d)$, where a is the class label, which is the accent in our case; and 3) the task decoder network G_y with parameters θ_y, such that $\boldsymbol{y} = G_y(\boldsymbol{f}; \theta_y)$, where $\boldsymbol{y}$ is the inferred transcription for an ASR model. Figure 1 shows this setup. The goal of DAT is to make the feature f generated by G_f to be domain-invariant (i.e., accent-invariant for our ASR task) for inputs from any accent $\boldsymbol{x} \in S$ or $\boldsymbol{x} \in U$, so that the decoder network G_y can be domain-agnostic and robust to unseen accents. For our QuartzNet model, we pick the intermediate layer C_3 for generating domain-invariant features.

The DAT objective function for a batch of N samples is:

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{a_i=0} \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \mathcal{L}_d^i(\theta_f, \theta_d) \quad (1)$$

For our ASR task with the QuartzNet model, $\mathcal{L}_y^i(\theta_f, \theta_y)$ is the CTC loss for decoding transcription characters, and $\mathcal{L}_d^i(\theta_f, \theta_d)$ is the cross-entropy (CE) loss for accent classification. Note that $\mathcal{L}_y^i(\theta_f, \theta_y)$ also has an indicator function as the coefficient, ensuring that CTC loss is only present for the annotated samples from the seen accent.

To optimize the parameters, we find the saddle points for the objective function as:

$$(\hat{\theta}_f, \hat{\theta}_y) = \underset{\theta_f, \theta_y}{\operatorname{argmin}} \, E(\theta_f, \theta_y, \hat{\theta}_d) \quad (2)$$

$$\hat{\theta}_d = \underset{\theta_d}{\operatorname{argmax}} \, E(\hat{\theta}_f, \hat{\theta}_y, \theta_d) \quad (3)$$

In DAT, this "min-max" optimization is done simultaneously within a single backward pass by using the gradient reversal layer (GRL) between feature encoder G_f and domain classifier G_d. The GRL layer acts as the identity function in the forward pass, but in the backward pass it multiplies the incoming gradients with $-\lambda$. Hence, the SGD update equations for DAT with a learning rate of μ can be written as follows:

$$\theta_f \longleftarrow \theta_f - \mu \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{a_i=0} \frac{\partial \mathcal{L}_y^i}{\partial \theta_f} - \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_f} \right) \quad (4)$$

$$\theta_y \longleftarrow \theta_y - \mu \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{a_i=0} \frac{\partial \mathcal{L}_y^i}{\partial \theta_y} \quad (5)$$

$$\theta_d \longleftarrow \theta_d - \mu \frac{1}{N} \sum_{i=1}^{N} \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_d} \quad (6)$$

3.3. Accent Pre-training for Robust DAT

In the DAT approach, the feature encoder and discriminator networks have competing objectives. Hence, during training, the feature encoder network is continuously learning from the gradients of the discriminator so as to generate domain-invariant features that would "fool" the discriminator. Nevertheless, if we start training with an untrained discriminator, it will send highly noisy signals back to the encoder, resulting in unstable training. A strategy was proposed to mitigate this by initializing $\lambda = 0$ and gradually increasing its value over the epochs [10]. We integrate this strategy in our experiments. However, we observe that when performing DAT combined with transfer learning, such fundamental parameter update strategies are not sufficient as we start with an already powerful encoder network. In this scenario, the discriminator is never able to "catch up" with the encoder, thus degrading the effect of adversarial training. This issue is exacerbated for transfer learning in the low-resource setting, when there is insufficient data to train a randomly initialized discriminator jointly with a fully tuned encoder.

In this work, we propose Accent Pre-training (Acc-PT) to overcome this issue, making DAT robust for low-resource transfer learning. We hypothesize that performing DAT transfer learning with a strong discriminator trained on the initial latent space leads to more stable training, since both encoder and discriminator are at an equal footing. In our Acc-PT approach, we initially freeze the feature encoder network and pre-train the discriminator until convergence. Once we have a pre-trained discriminator for the pre-trained ASR model, we then perform domain adversarial training. Table 2 outlines the pre-trained parameters for different training regimes. In our experiments, we empirically show that Acc-PT combined with DAT improves upon the performance of DAT for multiple accents.

Table 2: *Pre-trained parameters, "-" = not used.*

	θ_f	θ_y	θ_d
Transfer Learning	yes	yes	-
+ DAT	yes	yes	no
+ Acc-PT + DAT (**proposed**)	yes	yes	**yes**

4. Experiments

4.1. Data

We use Mozilla's Common Voice dataset [28] for our experiments. Common Voice is a crowd-sourced project that aims to collect natural human speech in different languages from a variety of demographics of people across the world. The source text for the speech is collected from various sources such as blog posts, movie dialogues etc. Since our QuartzNet model is pre-trained to recognize English speech, we use the English subset from Common Voice, which was reported to have 1.7k hours of speech at the time of access, with 1.3k hours of speech having been validated for correctness by majority voting. Contributors to this dataset can also self-report their speech accent which gets annotated for each speech sample. This is of most use to us since this allows us to extract accented speech with accent labels from the dataset. The source text is repeated multiple times across and within accents, making this dataset a good candidate for extracting accent-specific semantics.

We picked US accent as the standard seen accent as it has the largest amount of annotated speech (~255 hours). We chose

Table 3: *Details of different accents extracted from the validated samples of the Mozilla Common Voice dataset.*

Accent	# utterances	# hours
US	206,653	255.3
England	76,622	91.3
Indian	31,919	42.0
Australia	29,521	38.0
Scotland	8,405	11.9
African	5,430	7.0
Philippines	1,895	2.5
Total	360,445	448

Table 4: *Word error rates (WERs in %) for different approaches computed across multiple accents. Annotated speech from only the US accent is used while training with CTC, all other accents are treated as unannotated in our experiments. The numbers in parenthesis denote the number of utterances for each accent in the test set.*

	US (41,330)	*Unseen accents for CTC during training* England (15,321)	Indian (6,384)	Australia (5,904)	Scotland (1,681)	African (1,087)	Philippines (379)	Wt. Avg. *unseen* (30,756)	Wt. Avg. *all* (72,086)
Baseline	15.64	17.81	38.46	26.09	51.15	19.20	27.86	25.68	19.92
+ DAT	8.88	14.62	26.29	21.74	42.80	13.99	18.62	19.98	13.61
+ Acc-PT + DAT	8.92	14.77	26.26	21.99	43.10	14.28	18.23	20.11	13.70
Conventional CTC	8.98	13.88	25.96	21.20	42.14	13.92	18.31	19.39	13.42
+ DAT	8.84	14.03	**25.59**	21.30	42.32	13.87	18.42	19.42	13.35
+ Acc-PT + DAT	**8.81**	**14.01**	25.64	**20.71**	**42.08**	**13.85**	**18.09**	**19.29**	**13.28**

6 other accents as the target unseen accents, for which we only use the speech and ignore the annotations. We chose 3 accents (England, Indian, Australia) having relatively higher availability, and 3 accents (Scotland, African, Philippines) having very low availability in the dataset. For all 7 accents, we only use the validated speech samples, which total up to 448 hours. Note that all 7 accents combined is still less than half of 960 hours of LibriSpeech data used to train the QuartzNet model. Table 3 shows a summary of the all the accent data. For train, test and validation splits, we randomly partition the accented samples into 70%, 20% and 10% respectively.

4.2. Baseline

All ASR models in this work were trained using the NVIDIA NeMo toolkit [29]. We use the QuartzNet 15x5 model [4] that was pre-trained on LibriSpeech for our transfer learning baseline. Without any re-training, the baseline model has a micro average word error rate (WER) of 19.92% across all accents (Table 4). For DAT, we treat the output of block C_3 from the QuartzNet 15x5 model as the domain-invariant latent space. Hence, the QuartzNet blocks $\{C_1, B_1, \dots C_2, C_3\}$ can be seen as our feature encoder, and block C_4 as the task decoder. For the DAT discriminator input, we take the mean of the encoded sequence embeddings from the output of block C_3 as the generated feature $\boldsymbol{f}$. The discriminator architecture consists of a linear layer of output size 512 followed by 2 dense layers of output size 1024, finally ending with a linear layer which outputs the class scores with softmax acivation. Each dense layer consists of a linear layer, ReLu activation and dropout. For all re-training and DAT, we used the NovoGrad optimizer, which was used to originally train our baseline model. The network hyperparameters (e.g., λ) were tuned on the validation split.

4.3. Results

We summarize all results from our experiments in Table 4. The last 2 columns show the weighted average WER performance for the unseen accents as well as all accents combined.

Accent-invariant transcription. We see that even simple DAT with re-tuning the baseline model significantly boosts the performance across all accents. This can be attributed to the fact that there is a domain mismatch between the base task (LibriSpeech; more formal speech from audio books) and the new task (Common Voice; more informal speech from various sources such as movie dialogues). However, due to the limited amount of accented data available in the Common Voice dataset, Acc-PT does not improve upon the DAT performance from the baseline model even though we observed a high discriminator accuracy. More specifically, due to the domain mismatch between the base task and new task, the DAT training with an accent discriminator pre-trained on the base task introduces noise corresponding to the base task for accent regularization, thus driving the latent space away from the optimal target distribution for the new task.

Leveraging transfer learning for domain shift. We perform conventional CTC transfer learning by re-tuning the baseline model on the standard accent (US-English). We observe that it improves upon DAT performance since there is no longer any incoming signals from the discriminator that is domain specific to the base task. Comparing transfer learning performance of DAT and Acc-PT+DAT on the conventional CTC model reveals that Acc-PT is now able to better regularize the latent space to make it accent invariant with respect to the optimal target distribution for the new task. Our Acc-PT+DAT approach shows the best WER across 5 out of 6 unseen accents.

Additionally, we observe that simple DAT in this case improves overall WER (from 13.42% to 13.35%) mainly due to improvement in US accent, whereas performance on unseen accents actually degrades with respect to conventional CTC (from 19.39% to 19.42%). Noise backpropagated from an untrained, randomly initialized discriminator during initial epochs leads to sub-optimal adversarial training. In contrast, Acc-PT with a pre-trained discriminator leads to better DAT training with transfer learning, even improving the US-accent performance along with having the best unseen accent performance.

We also note here that the magnitude of improvement of our Acc-PT+DAT approach over the other methods (e.g., from 19.39% to 19.29%) could be much more pronounced for ASR models that are already not as powerful as the QuartzNet 15x5 model that we use in this study. We intentionally choose a highly robust baseline so that our results are beneficial for practitioners interested in deploying real-world ASR systems.

5. Conclusion

In this work, we explore domain adversarial training (DAT) combined with transfer learning to improve the accent robustness of a trained ASR model in the low-resource setting. We propose the Accent Pre-training (Acc-PT) strategy of pre-training the discriminator on unannotated accented speech samples when performing DAT with a trained model. Our experiments show that Acc-PT with DAT improves the ASR performance by 33% on average across multiple accents, when annotated samples from only a single accent were available.

6. References

[1] D. Yu and L. Deng, *AUTOMATIC SPEECH RECOGNITION*. Springer, 2016.

[2] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[3] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6074–6078.

[4] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.

[5] M. Sunkara, S. Ronanki, D. Bekal, S. Bodapati, and K. Kirchhoff, "Multimodal Semi-Supervised Learning Framework for Punctuation Prediction in Conversational Speech," in *Proc. Interspeech 2020*, 2020, pp. 4911–4915. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3074

[6] A. Shenoy, S. Bodapati, M. Sunkara, S. Ronanki, and K. Kirchhoff, "Adapting long context nlm for asr rescoring in conversational agents," 2021.

[7] G. Synnaeve, Q. Xu, J. Kahn, E. Grave, T. Likhomanenko, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[8] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, and J. Černockỳ, "Semi-supervised sequence-to-sequence asr using unpaired speech and text," *arXiv preprint arXiv:1905.01152*, 2019.

[9] Y. Long, Y. Li, S. Wei, Q. Zhang, and C. Yang, "Large-scale semi-supervised training in deep learning acoustic model for asr," *IEEE Access*, vol. 7, pp. 133 615–133 627, 2019.

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] 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)*. IEEE, 2018, pp. 4854–4858.

[12] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[13] A. Tripathi, A. Mohan, S. Anand, and M. Singh, "Adversarial learning of raw speech features for domain invariant speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5959–5963.

[14] K. Hu, H. Sak, and H. Liao, Eds., *Adversarial Training for Multilingual Acoustic Modeling*, 2019. [Online]. Available: https://arxiv.org/pdf/1906.07093.pdf

[15] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition." in *Interspeech*. San Francisco, CA, USA, 2016, pp. 2369–2372.

[16] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant representations for noisy speech recognition," *arXiv preprint arXiv:1612.01928*, 2016.

[17] C. Luu, P. Bell, and S. Renals, "Channel adversarial training for speaker verification and diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7094–7098.

[18] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6216–6220.

[19] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning for speaker verification." in *Interspeech*, 2019, pp. 4315–4319.

[20] K. Rao and H. Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4815–4819.

[21] S. Ghorbani and J. H. Hansen, "Leveraging native language information for improved accented speech recognition," *Proc. Interspeech 2018*, pp. 2449–2453, 2018.

[22] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in *INTERSPEECH*, 2018, pp. 2454–2458.

[23] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *INTERSPEECH*, 2019, pp. 2140–2144.

[24] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *Proc. Interspeech 2019*, pp. 71–75, 2019.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[27] B. Ginsburg, P. Castonguay, O. Hrinchuk, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, H. Nguyen, Y. Zhang, and J. M. Cohen, "Stochastic gradient methods with layer-wise adaptive moments for training of deep networks," *arXiv preprint arXiv:1905.11286*, 2019.

[28] "Mozilla Common Voice," https://commonvoice.mozilla.org.

[29] "NVIDIA NeMo," https://github.com/NVIDIA/NeMo.