REDAT: ACCENT-INVARIENT REPRESENTATION FOR END-TO-END ASR BY DOMAIN ADVERSARIAL TRAINING WITH RELABELING

Hu Hu*, Xuesong Yang†, Zeynab Raeesy†, Jinxi Guo†, Gokce Keskin†, Harish Arsikere†, Ariya Rastrow†, Andreas Stolcke†, Roland Maas†

*Georgia Institute of Technology, Atlanta, USA †Amazon.com, Seattle, USA

ABSTRACT

Accents mismatching is a critical problem for end-to-end ASR. This paper aims to address this problem by building an accent-robust RNN-T system with domain adversarial training (DAT). We unveil the magic behind DAT and provide, for the first time, a theoretical guarantee that DAT learns accent-invariant representations. We also prove that performing the gradient reversal in DAT is equivalent to minimizing the Jensen-Shannon divergence between domain output distributions. Motivated by the proof of equivalence, we introduce reDAT, a novel technique based on DAT, which relabels data using either unsupervised clustering or soft labels. Experiments on 23K hours of multi-accent data show that DAT achieves competitive results over accent-specific baselines on both native and non-native English accents but up to 13% relative WER reduction on unseen accents; our reDAT yields further improvements over DAT by 3% and 8% relatively on non-native accents of American and British English.

Index Terms—Accent-invariance, end-to-end ASR, domain adversarial training, multi-accent ASR, RNN transducer

1. INTRODUCTION

Recent application of recurrent neural network transducers (RNN-T) has achieved significant progress in the area of online streaming end-to-end automatic speech recognition (ASR) [1–4]. However, building an accent-robust system remains a big challenge. Accents represent systematic variations within a language as a function of geographical region (e.g. British versus American English), social group, or other factors such as nativeness of speakers. Accents occur in many gradations and commercial speech applications typically only model varieties associated with major countries. For example in real-world smart speaker devices, users set up their language preferences regardless of whether they are native speakers or not; thus ASR systems trained mainly on only native speech risk degradation when faced with non-native speech.

Accent-robust ASR systems aim to mitigate the negative effects of non-native speech and recent works have made some progress. A straightforward exploration is to build an accent-specific system where accent information, such as i-vectors, accent IDs, or accent embeddings, are explicitly fed into the neural networks along with acoustic features [5–10]. These approaches typically either adapt a unified model with accent-specific data, or build a separate decoder for each accent. These accent-specific models perform well on their accent-specific test sets, but they do not generalize well to unseen accents. Accent-invariant systems [11, 12], on the other hand, aim to build a universal model that learns accent-invariant features that can be expected to generalize to new accents. For example, simply pooling data across all accents during training brings in additional variations so that the models are capable of learning accent-invariant information. The adversarial training methods [13,14] also help to achieve the same goal through the gradient reversal.

We aim to advance accent-invariant modeling with RNN-T based on the domain adversarial training (DAT). DAT is expected to learn accent-invariant features by reversing gradients propagated from the accent classifier. Our experiments demonstrate DAT can achieve competitive performance on native, non-native, and unseen accents. This paper makes the following novel contributions:

• We lay out the theory behind DAT and we provide, for the first time, a theoretical guarantee that DAT learns accent-invariant representations.

• We also prove that performing the gradient reversal in DAT is equivalent to minimizing the Jensen-Shannon divergence between output distributions from different domain classes.

• Motivated by the proof of equivalence, we introduce reDAT, a novel technique based on DAT, which refines accent classes with either unsupervised clustering or soft labels. reDAT yields significant improvements over strong baselines on non-native and unseen accents without sacrificing of native speech performance.

2. DAT FOR ACCENTED SPEECH RECOGNITION
Train features. the generator gradient values of $R\alpha$ where $\alpha$ is the learning rate and $\lambda$ is the linear weight of two different losses are accessed. $L_C$ is the CE loss of accent classifier outputs, and it passes the negative gradient to $G$ so that the ability to distinguish accents in the generator outputs is minimized. In other words, the output $z$ from the generator $G$ is expected to embed accent-invariant features.

The weight matrices of $G$, $C$, $R$ are denoted as $\theta_G$, $\theta_C$, $\theta_R$. The losses are denoted as $L_G$, $L_C$, $L_R$. Each weight is updated by the following gradient descent rules,

$$\theta_G \leftarrow \theta_G - \alpha \left( \frac{\partial L_R}{\partial \theta_G} - \lambda \frac{\partial L_C}{\partial \theta_G} \right),$$

$$\theta_C \leftarrow \theta_C - \alpha \frac{\partial L_C}{\partial \theta_C},$$

$$\theta_R \leftarrow \theta_R - \alpha \frac{\partial L_R}{\partial \theta_R},$$

where $\alpha$ is the learning rate and $\lambda$ is the linear weight of two gradient values of $R$ and $C$. We freeze $G$ and $R$ to perform forward inference.

### 2.1. Theoretical Guarantee of DAT for Accent-Invariance

DAT methods have been widely used in robust ASR systems, however, there is little theoretical analysis to explain why DAT methods are capable of learning expressive domain-invariant features. We extend the theory behind generative adversarial networks (GAN) [19] and proved that performing gradient reversal is equivalent to minimizing Jensen-Shannon divergence among output domain distributions. This finding generalizes to any domain mismatch problems. Here we focus on the accent-invariance problems.

The output distribution of the $i$-th accent from $G$ is denoted as $P_{Gi}$, where $i \in [1,N]$ is an accent index and $N$ is the number of accents. Given that $G$ is fixed during training, we could obtain the optimal $C^*$ by minimizing their cross-entropy or equivalently maximizing the log-likelihood as,

$$C^* = \arg \max_{\theta_C} \sum_i^{N} E_{x \sim P_{Gi}(x)} \log C_i(z).$$

(1)

Softmax is used as the final output layer so that $C_i(z)$ is the probability of an input utterance belonging to each accent. So we have the following constraints,

$$\sum_i^{N} C_i(z) = 1, \quad 0 < C_i(z) < 1.$$  

(2)

Eq (1) and Eq (2) indicate $C^*$ is convex and has a global maximum since the 2nd-order derivative of every variable $C_i(z)$ is always negative. This optimization problem could be solved by linear programming as,

$$C^*_i(z) = \frac{P_{Gi}}{\sum_i^{N} P_{Gi}}.$$  

(3)

Our solution is similar to GAN’s [19] but extends to multiple variables. The generator $G$ connects two tasks so that two different losses are accessed. $L_C$ is the CE loss of accent classifier outputs, and it passes the negative gradient to $G$. If we only consider the effect of $L_C$ on $G$, we have the optimal $G^*$ as,

$$G^* = \arg \min_{\theta_G} \left( \arg \max_{\theta_C} \sum_i^{N} E_{x \sim P_{data}(x)} \log C_i(G(x)) \right).$$  

(4)

During the parameter updating of $G^*$, $C$ is fixed. We can obtain the solution of $C^*_i(z)$ by plugging Eq (3) into Eq (4). After deduction and simplification, we have the optimal $G^*$ as,

$$G^* = \arg \min_{\theta_G} \left( - \log N + \sum_i^{N} KLD \left( P_{Gi} \parallel \frac{\sum_i^{N} P_{Gi}}{N} \right) \right),$$

where $KLD$ is the Kullback–Leibler divergence. It is equivalent to minimize the JSD between the distributions of all accents:

$$G^* = \arg \min_{\theta_G} \left( - \log N + JSD \left( P_{G1}, P_{G2}, \ldots, P_{GN} \right) \right).$$

Hence, we have shown that performing gradient reversal is equivalent to minimizing Jensen-Shannon divergence between output distributions from different accents. The global minimum is achieved if and only if $P_{G1}=P_{G2}=\ldots=P_{GN}$, which indicates that the embeddings $z$ are accent-invariant.

### 3. REDAT: DAT WITH RELABELING

The theoretical proof of the equivalence between gradient reversal and minimizing JSD of output distributions from accents suggests that we should get more invariant training results by predefining more detailed acoustic information, such as different noise conditions 'unsup', and multiple accents and languages [14-18]. The proposed DAT training framework consists of an accent-invariant feature generator $G$, English accent classifier $C$, and RNN-T model $R$ (see Fig 1). We choose LSTM layers for both the feature generator and accent classifier, and the RNN-T model includes encoder, decoder, and joint networks. In the training phase, negative gradients (blue arrow) are back-propagated from the generator to the accent classifier so that the ability to distinguish accents in the generator outputs is minimized. In other words, the output $z$ from the generator $G$ is expected to embed accent-invariant features.

![Fig. 1. REDAT framework by relabeling with either unsupervised clustering ('unsup') or soft labels ('soft').](image-url)

The output distribution of the $i$-th accent from $G$ is denoted as $P_{Gi}$, where $i \in [1,N]$ is an accent index and $N$ is the number of accents. Given that $G$ is fixed during training, we could obtain the optimal $C^*$ by minimizing their cross-entropy or equivalently maximizing the log-likelihood as,

$$C^* = \arg \max_{\theta_C} \sum_i^{N} E_{x \sim P_{Gi}(x)} \log C_i(z).$$

(1)

Softmax is used as the final output layer so that $C_i(z)$ is the probability of an input utterance belonging to each accent. So we have the following constraints,

$$\sum_i^{N} C_i(z) = 1, \quad 0 < C_i(z) < 1.$$  

(2)

Eq (1) and Eq (2) indicate $C^*$ is convex and has a global maximum since the 2nd-order derivative of every variable $C_i(z)$ is always negative. This optimization problem could be solved by linear programming as,

$$C^*_i(z) = \frac{P_{Gi}}{\sum_i^{N} P_{Gi}}.$$  

(3)

Our solution is similar to GAN’s [19] but extends to multiple variables. The generator $G$ connects two tasks so that two different losses are accessed. $L_C$ is the CE loss of accent classifier outputs, and it passes the negative gradient to $G$. If we only consider the effect of $L_C$ on $G$, we have the optimal $G^*$ as,

$$G^* = \arg \min_{\theta_G} \left( \arg \max_{\theta_C} \sum_i^{N} E_{x \sim P_{data}(x)} \log C_i(G(x)) \right).$$  

(4)

During the parameter updating of $G^*$, $C$ is fixed. We can obtain the solution of $C^*_i(z)$ by plugging Eq (3) into Eq (4). After deduction and simplification, we have the optimal $G^*$ as,

$$G^* = \arg \min_{\theta_G} \left( - \log N + \sum_i^{N} KLD \left( P_{Gi} \parallel \frac{\sum_i^{N} P_{Gi}}{N} \right) \right),$$

where $KLD$ is the Kullback–Leibler divergence. It is equivalent to minimize the JSD between the distributions of all accents:

$$G^* = \arg \min_{\theta_G} \left( - \log N + JSD \left( P_{G1}, P_{G2}, \ldots, P_{GN} \right) \right).$$

Hence, we have shown that performing gradient reversal is equivalent to minimizing Jensen-Shannon divergence between output distributions from different accents. The global minimum is achieved if and only if $P_{G1}=P_{G2}=\ldots=P_{GN}$, which indicates that the embeddings $z$ are accent-invariant.

### 3. REDAT: DAT WITH RELABELING

The theoretical proof of the equivalence between gradient reversal and minimizing JSD of output distributions from accents suggests that we should get more invariant training results by predefining more detailed acoustic information, such
as a refined accent label for each utterance. In addition, taking into account that the boundary between accents are very fuzzy and there are non-native accent utterances in most accent-specific data sets, we further propose reDAT by introducing relabeling approaches for domain classes in order to refine the labeling. Two relabeling approaches, unsupervised clustering and soft labels, are investigated.

3.1. Relabeling with Unsupervised Clustering

We relabel utterance accents in an unsupervised manner (‘unsup’) using a three-phase processes as shown in Figure 1. An utterance-level accent classifier is trained with original accent labels. Then we use this well-trained accent classifier to extract utterance-level embeddings, which are expected to contain detailed and distinct accent information. Lastly, we predict new domain labels for utterances by performing the unsupervised clustering on this set of utterance embeddings. We specify a number of clusters larger than the original number of accents so that the new labels encode additional detail. DAT could directly benefit from the newly generated accent domain labels and improve its generalization ability to non-native and unseen accents. In this paper we choose K-means for the clustering phase.

3.2. Relabeling with Soft Labels

We can also refine original accent labels with soft labels (‘soft’) in a two-phase processes as shown in Figure 1. An utterance-level accent classifier is trained with original accent labels. Then we generate a soft label for each utterance from this accent classifier. DAT is performed based on these newly generated soft labels. Previous studies have found that soft labels correlate with structural relationship among accents [20, 21], so that we expect them to encode more detailed accent information. Noted that although one-hot labels are replaced by soft labels, the theoretical equivalence to minimizing JSD still holds. Performing gradient reversal on soft labels is also equivalent to minimizing the JSD, but it is the JSD between each utterance distribution. When one-hot labels are replaced by soft labels, the expression of $C^*$ in Eq (1) is replaced as,

$$C^* = \arg \max_{G} E_{z \sim P_G(z)} \sum_{i} l_i(x) \log G_i(z),$$

where $l_i(x)$ is a scalar, indicating the soft label for each utterance, which depends on the input $x$ and trained accent classifier $l_i$. Then $G^*$ is derived as,

$$G^* = \arg \min_{G} (-N \log N + \text{JSD}(l(x_1) \cdot P_G, \ldots, l(x_N) \cdot P_G)),$$

where $l(x_i) \cdot P_G$ is a distribution depending on the input $x$, which can be regarded as the linear combination of different accent distributions. Thus, by using soft labels for gradient reversal, we replace minimizing JSD between accent distributions with doing the same between utterance distributions.

4. EXPERIMENTS

4.1. Experimental Setup

For our experiments we used de-identified human labelled speech data (23K hours) from voice controlled far-field and close-talk devices. This data set consists of English recordings from 3 different regions, including 13K hours of en-US data, 6K hours of en-GB data, and 4K hours of en-IN (Indian English) data. Each utterance in the en-US and en-GB test sets has a label that characterizes the speaker as native or non-native. Most of the recordings (over 90%) are from native speakers. In addition, to evaluate generalization, we use extra en-AU (Australian English) data as an unseen test set. Since the en-IN data lacks nativeness labels and is smaller in size, we only evaluate on en-US, en-GB, and en-AU test sets.

All experiments use 64-dimensional log-Mel features, computed over 25ms windows with 10ms hop length. Each feature vector is stacked with 2 frames to the left and down-sampled to a 30ms frame rate. All experiments are performed with an RNN-T model. The baseline RNN-T model consists of an encoder, a prediction network, and a joint network. The encoder consists of 3 LSTM layers with the hidden dimension of 1024, whereas the prediction network consists of 2 LSTM layers with the hidden dimension of 1024 and the embedding size of 512. We adopt a simple addition strategy in the joint network to combine outputs from encoder and prediction networks to limit memory and computation. The softmax layer consists of 10K output units and is trained to predict word-piece tokens, which are generated using the byte pair encoding algorithm [22]. To apply the reDAT framework to the RNN-T model, the first two LSTM encoder layers of RNN-T serve as the generator, whose outputs are fed into a domain classifier as well as into the remaining parts of the RNN-T encoder. The accent classifier consists of 2 LSTM layers with the hidden dimension of 1024 and predicts three accent classes, i.e. en-US, en-GB, and en-IN.

All models are trained using the Adam optimizer [23], with a learning rate schedule including an initial linear warm-up phase, a constant phase, and an exponential decay phase [4]. All the baseline models and proposed methods use the same training strategy. Specifically, the learning rates for the constant phase and end of the exponential decay phase are $5e^{-4}$ and $1e^{-5}$, respectively. During the training stage, the acoustic training data is augmented with the SpecAugment [24] to improve the robustness.

4.2. Baselines

For comparison purposes, we also investigate recent popular multi-accent speech recognition approaches. Two accent-invariant approaches and two accent-specific approaches are compared. The results of baseline systems are shown in the 2nd to 5th rows in Table 1. Data pooling refers to the simple data pooling strategy, which combines all data together...
Table 1. Normalized WERs on 23K hours of en-X data. AS or AI denotes an accent-specific or accent-invariant model; native or non-native denotes native or non-native speakers on test sets; unsup8 or unsup20 denotes reDAT with 8 or 20 unsupervised clusters; soft denotes reDAT with soft labels.

| Approach     | AS/AI | en-US % | en-GB % | en-AU % |
|--------------|-------|---------|---------|---------|
|              |       | native  | non-native | avg. | native  | non-native | avg. | (unseen) |
| Data pooling | AI    | 1.000   | 1.472   | 1.027 | 1.315   | 1.574   | 1.315 | 1.393   |
| AIPNet-s     | AI    | 0.997   | 1.425   | 1.023 | 1.330   | 1.543   | 1.332 | 1.412   |
| One-hot embedding | AS | 0.981   | 1.528   | 1.010 | 1.284   | 1.540   | 1.284 | 1.574   |
| Linear embedding | AS | 0.991   | 1.442   | 1.017 | 1.284   | 1.534   | 1.282 | 1.569   |
| DAT          | AI    | 0.985   | 1.448   | 1.012 | 1.293   | 1.567   | 1.294 | 1.373   |
| reDAT-unsup8 | AI    | 0.969   | 1.472   | 0.996 | 1.270   | 1.465   | 1.266 | 1.359   |
| reDAT-unsup20 | AI    | 0.980   | 1.470   | 1.006 | 1.282   | 1.492   | 1.280 | 1.361   |
| reDAT-soft   | AI    | 0.973   | 1.409   | 0.997 | 1.309   | 1.440   | 1.307 | 1.388   |

and trains a unified model. One-hot embedding and linear embedding refer to the accent-specific approaches which utilize external accent information. Specifically, one-hot or linear accent embeddings are added as inputs to all layers of the RNN-T model. One-hot embeddings directly use one-hot accent labels whereas linear embedding uses an extra linear transformation matrix to map one-hot labels into linear embedding vectors. For those accent-specific systems, accent information is required in both the training and evaluation phase. Thus, intuitively, this approach cannot handle unseen accents as well, since no matched data was observed in training. AIPNet-s refers to the recently proposed AIPNet system. An extra accent-invariant GAN and decoder layer are introduced for pre-training. Then the ASR model and invariant feature generator are trained jointly. We simplify the original AIPNet framework by replacing accent-specific GAN with one-hot labels, which can also provide the accent-specific information.

4.3. Experimental Results on 23K Hours of en-X Data

The experimental results of the normalized word error rates (WERs) are shown in Table 1. At first, when comparing the results on native and non-native speakers, we can see that although we may achieve good ASR performance on native speakers, there is still a big performance gap between native speakers and non-native speakers. The 2nd to 5th rows of Table 1 show different multi-accent ASR methods as baselines. As for accent-invariant baselines, AIPNet-s can bring gains on non-native data over data pooling. As for accent-specific approaches, i.e., one-hot embeddings and linear embeddings, they show better performance than data pooling on native data, but do not generalize well to the unseen accent test set. That is because these two models do not know the accent label for the en-AU test set, and they have to choose another accent-specific model (e.g., en-US), even though the accents are mismatched.

Experimental results of DAT and our proposed reDAT are shown in the last four rows of Table 1. When compared to AS and AI baselines, DAT achieves competitive WERs on both native and non-native accents but up to 13% relative reduction on unseen accents; the best performance of reDAT with 8 unsupervised clusters shows relative WER reductions of 2% to 4% over the data pooling baseline and 2% over DAT, respectively. When increased to 20 unsupervised clusters, we observe a WER degradation over 8 clusters. On non-native accents, our reDAT with soft labels achieves significant improvements over DAT by 3% on en-US and 8% on en-GB, and over the best AS and AI baselines by 1% on en-US and 6% on en-GB. On native and unseen accents, we observe that reDAT with soft labels has very competitive results over original DAT.

5. CONCLUSION

This paper suggests a feasible solution to address the accents mismatching problem for end-to-end RNN-T ASR using DAT. We demonstrate that DAT could achieve competitive WERs over accent-specific baselines on both native and non-native English accents but significantly better WERs on unseen accents. We provide, for the first time, a theoretical guarantee that DAT extracts accent-invariant representations that generalize well across accents, and also prove that performing gradient reversal in DAT is equivalent to minimizing JSD between domain distributions. The proof of equivalence further motivates to introduce a novel method reDAT that yields relative WERs over DAT on non-native accents by a large margin.

1Normalized WER of a control model is calculated as the WER percentage over the reference. For example, Data Pooling is chosen as the reference so that its WER is 1.000, and DAT, as a control, is 0.985.
References

[1] Kanishka Rao, Hasım Sak, and Rohit Prabhavalkar, “Exploring architectures, data and units for streaming end-to-end speech recognition with rn-transducer,” in Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017, pp. 193–199.

[2] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kanan, Yonghui Wu, Ruoming Pang, et al., “Streaming end-to-end speech recognition for mobile devices,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6381–6385.

[3] Jinyu Li, Rui Zhao, Hu Hu, and Yifan Gong, “Improving rn-transducer modeling for end-to-end speech recognition,” in Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 114–121.

[4] Jinx Guo, Gautam Tiwari, Jasha Droppo, Maarten Van Segbroeck, Che-Wei Huang, Andreas Stolcke, and Roland Maas, “Efficient minimum word error rate training of rn-transducer for end-to-end speech recognition,” arXiv preprint arXiv:2007.13802, 2020.

[5] Thibault Viglino, Petr Motlicek, and Milos Cernak, “End-to-end accented speech recognition,” in INTERSPEECH, 2019, pp. 2140–2144.

[6] Kanishka Rao and Hasım Sak, “Multi-accent speech recognition with hierarchical grapheme based models,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 4815–4819.

[7] Sanghyun Yoo, Incheul Song, and Yoshua Bengio, “A highly adaptive acoustic model for accurate multi-dialect speech recognition,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5716–5720.

[8] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, “Joint modeling of accents and acoustics for multi-accent speech recognition,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 1–5.

[9] Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zifeng Chen, Yanghui Wu, and Kanishka Rao, “Multi-dialect speech recognition with a single sequence-to-sequence model,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4749–4753.

[10] Zhong Meng, Hu Hu, Jinyu Li, Changliang Liu, Yan Huang, Yifan Gong, and Chin-Hui Lee, “L-Vector: Neural label embedding for domain adaptation,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7389–7393.

[11] Dimitra Vergyi, Lori Lamel, and Jean-Luc Gauvain, “Automatic speech recognition of multiple accented English data,” in INTERSPEECH, 2010.

[12] Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro Moreno, and Austin Waters, “Towards acoustic model unification across dialects,” in Spoken Language Technology Workshop (SLT). IEEE, 2016, pp. 624–628.

[13] Yi-Chen Chen, Zhaojun Yang, Ching-Feng Yeh, Mahaveer Jain, and Michael I. Seltzer, “AIPNET: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6979–6983.

[14] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Osten- dorf, and Lei Xie, “Domain adversarial training for accented speech recognition,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4854–4858.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, “Domain-adversarial training of neural networks,” The Journal of Machine Learning Research, vol. 17, no. 1, pp. 2096–2030, 2016.

[16] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang, “Speaker-invariant training via adversarial learning,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5969–5973.

[17] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi, “Data decisions and theoretical implications when adversarially learning fair representations,” arXiv preprint arXiv:1707.00075, 2017.

[18] Ke Hu, Hasan Sak, and Hank Liao, “Adversarial training for multilingual acoustic modeling,” arXiv preprint arXiv:1906.07093, 2019.

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial networks,” 2014.

[20] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, “Relational knowledge distillation,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 3962–3971.

[21] Hu Hu, Sabato Marco Siniscalchi, Yannan Wang, and Chin-Hui Lee, “Relational teacher student learning with neural label embedding for device adaptation in acoustic scene classification,” arXiv preprint arXiv:2008.00110, 2020.

[22] Mike Schuster and Kaisuke Nakajima, “Japanese and Korean voice search,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 5149–5152.

[23] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[24] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv:1904.08779, 2019.