reDAT: Accent-Invariant Representation for End-to-End ASR by Domain Adversarial Training with Relabeling

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

*Georgia Institute of Technology
†Amazon Alexa
Overview

- Introduction
- Domain adversarial training (DAT)
- Theoretical Guarantee of Accent-Invariance for DAT
- reDAT: DAT with relabeling
  - Relabeling with unsupervised clustering
  - Relabeling with soft labels
- Experiments
  - Experimental setup
  - Experimental results and nativity analysis
- Conclusions
• **Accent mismatching is a critical problem for ASR**
  - Commercial speech applications typically only model varieties associated with major countries.
  - In real-world smart speaker devices, users set up their language preferences regardless of whether they are native speakers or not.
  - Therefore, ASR systems trained mainly on only native speech risk degradation when faced with non-native speech.

• **Previous works**
  - Accent-specific approaches: i-vectors, accent IDs, accent embeddings, ...
  - Accent-invariant approaches: data pooling, adversarial training, ...

• **Our goals**
  - Build an accent-invariant end-to-end ASR model robust to many English accents (e.g. en-US/en-GB/en-IN/en-AU/...).
  - Improve the recognition accuracy on native, non-native, and even unseen accents.
We aim to advance accent-invariant modeling with RNN-T based on the domain adversarial training (DAT).

Our 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 prove that performing the gradient reversal in DAT is equivalent to minimizing the Jensen-Shannon divergence.
- Motivated by the proof of equivalence, we introduce reDAT, a novel technique based on DAT.
- Results show significant improvements over strong baselines.
Domain Adversarial Training

- Domain adversarial training\(^1\) (DAT) is the basic framework of gradient reversal based methods.
- It uses an extra Accent Classifier to learn the accent invariant features.
- With the negative gradient back-propagated from accent classifier, Generator tends to be unable to distinguish the accent classes. This pushes the output embedding \(z\) to be invariant to accent.
- Model outputs:
  - RNN-T output -- positive gradient.
  - Accent Classifier output -- negative gradient.

Domain Adversarial Training (cont’d)

- **Training Stage**: the whole network is updated based on the following gradient decent rules:
  - **Generator**:
    \[
    \theta_G \leftarrow \theta_G - \alpha \left( \frac{\partial \mathcal{L}_R}{\partial \theta_G} - \lambda \frac{\partial \mathcal{L}_C}{\partial \theta_G} \right)
    \]
  - **RNN-T**:
    \[
    \theta_R \leftarrow \theta_R - \alpha \frac{\partial \mathcal{L}_C}{\theta_R}
    \]
  - **Accent Classifier**:
    \[
    \theta_C \leftarrow \theta_C - \alpha \frac{\partial \mathcal{L}_R}{\theta_C}
    \]
  - The accent IDs are needed during training, but not during inference.

---

**DAT framework**

- **Accent Classifier (C)**
- **RNN-T (R)**
- **Generator (G)**

CE loss

RNN-T loss

Reverse gradient

\[x\]

\[z\]
Theoretical Guarantee of Accent-Invariance for DAT

- **Claim:** Performing gradient reversal in DAT is equivalent to minimizing the Jensen-Shannon divergence (JSD) between output distributions from different accents.

- **Proof:**
  - For accent classifier $C$, we can find optimal $C^*$ by minimizing the CE loss,
    \[ C^* = \arg \max_{\theta_C} \sum_{i} E_{z \sim P_{G_i}} \log C_i(z) \]
  - **Softmax** is applied s.t. the following constraint holds,
    \[ \sum_{i} C_i(z) = 1, \quad 0 < C_i(z) < 1 \]
  - $C^*$ is convex since 2nd-order derivative of $C_i(z)$ is negative. We can find the only solution by linear programing,
    \[ C^*_i(z) = \frac{P_{G_i}}{\sum_{i}^N P_{G_i}} \]
  - For generator $G$, we can find optimal $G^*$ if we only consider CE loss onto $G$,
    \[ G^* = \arg \min_{\theta_G} \left( \arg \max_{\theta_C} \sum_{i} E_{z \sim P_{\text{data}}} \log C_i(G(z)) \right) \]
  - Hence, for $G$, it can be deduced to \( G^* = \arg \min_{G} \left( -N \log N + JSD (P_{G_1}, P_{G_2}, \ldots, P_{G_N}) \right) \)
• **From the theory proof:**
  - Gradient reversal is equivalent to minimize the JSD between output distributions from different classes.
  - The global minima is achieved iff. \( P_{G_1} = P_{G_2} = \ldots = P_{G_N} \), which indicates that the embeddings \( z \) are accent-invariant.
  - We should get better results by predefining more detailed acoustic information, i.e., more accurate accent labels.

• **Proposed relabeling approaches:**
  - Perform unsupervised clustering to get more accurate accent labels.
  - Use soft labels for gradient reversal instead of hard labels.
Relabeling with Unsupervised Clustering

We should get better results by predefining more fine-grained accent labels motivated by the proof of equivalence.

Procedure:

<Step 1>: Train an accent classifier in a supervised way on our en-X datasets.

<Step 2>: Generate utterance-level accent embeddings by the accent classifier.

<Step 3>: Perform unsupervised clustering on accent embedding. We use K-means.

<Step 4>: Perform DAT on new labels. The number of new labels is equal to the number of clusters.
Relabeling with Unsupervised Clustering (cont’d)

- Visualization of unsupervised clustering on generated embedding (Step 3) by t-SNE:

  ![Visualization](image)

  **Target accent distribution**
  (en-US, en-GB, en-IN)

  **K-means clustering distribution**
  (8 classes)

- 3 classes blur accent boundaries. Many utterances are in the overlapping regions where classes are hard to discriminate.
- The fine-grained 8 classes are capable of capturing more detailed and non-native English accents.
Relabeling with Soft Labels

Compared with one-hot labels, soft labels are expected to do a better job of representing accents compared to one-hot vector owing to the fuzziness of accent boundaries.

Procedure:

<Step 1>: Train an accent classifier in a supervised way on our en-X datasets.

<Step 2>: Generate soft labels for each utterance by the accent classifier.

<Step 3>: Finally perform DAT on new generated soft labels.
• Although one-hot labels are replaced by soft labels, the theory guarantee still holds. It’s still equivalent to minimize the JSD, but between different distributions.

• Optimal $C$:
  • One-hot labels: $C^* = \arg \max_{\theta_C} \sum_i^N E_{z \sim P_{G_i}(z)} \log C_i(z)$
  • Soft labels: $C^* = \arg \max_{\theta_C} E_{z \sim P_G(z)} \sum_i^N l_i(x) \log C_i(z)$

• Optimal $G$:
  • One-hot labels: $G^* = \arg \min_G \left( -N \log N + \text{JSD} (P_{G_1}, P_{G_2}, \ldots, P_{G_N}) \right)$
  • Soft labels: $G^* = \arg \min_{\theta_G} \left( -N \log N + \text{JSD} (l(x_1) \cdot P_G, \ldots, l(x_N) \cdot P_G) \right)$

• By using soft labels for gradient reversal, we move from minimizing JSD between output distributions to minimizing JSD between each utterance distribution.
Experimental Setup

• **Data sets:**
  - ~23K hours en-X data in total
    - ~13K hours en-US
    - ~6K hours en-GB
    - ~4K hours en-IN
  - Extra en-AU data is used as unseen test set.

• **Model:**
  - Our experiments are based on an RNN-T model.
    - 5-layer 1024 LSTM as encoder.
    - 2-layer 1024 LSTM as prediction network.
  - 10K word-pieces as target tokens.

• **Training and evaluation:**
  - Spectral augmentation is used for training.
  - We pool all accent data with sampling probability in proportion to accent-specific corpus size, and train a unified model.
  - Beam search with a size of 16 is used for decoding.
Experiments

- **Baseline approaches:**
  - Data pooling (M0):
    - Combines data of all accents and trains a unified model.
  - One-hot embeddings\(^{[1]}\) (M1):
    - Append one-hot accent labels to the outputs of each layer in the RNN-T model.
  - Linear embeddings\(^{[1]}\) (M2):
    - Based on one-hot embedding, a transform matrix is utilized to map one-hot labels into linear embedding vectors.
  - AIPNet-s\(^{[2]}\) (M3):
    - An extra accent-invariant GAN and decoder layer are introduced for pre-training and jointly trains ASR model and invariant feature generator.
    - We simplified it as AIPNet-s by replacing accent-specific GAN with accent labels.

- Data-pooling and AIPNet-s are accent-invariant (AI) systems, where accent information is not required in the evaluation stage.
- One-hot embeddings and linear embeddings are accent-specific (AS) systems, where accent information is required in the evaluation stage.

\(^{[1]}\) Li, Bo, et al. "Multi-dialect speech recognition with a single sequence-to-sequence model." In ICASSP'18.

Results

- Results on 23K hour en-X data (normalized WER$^1$)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AS/AI</th>
<th>en-US</th>
<th>en-GB</th>
<th>en-AU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>native</td>
<td>non-native</td>
<td>avg.</td>
</tr>
<tr>
<td>M0: Data pooling</td>
<td>AI</td>
<td>1.000</td>
<td>1.472</td>
<td>1.027</td>
</tr>
<tr>
<td>M1: AIPNet-s</td>
<td>AI</td>
<td>0.997</td>
<td>1.425</td>
<td>1.023</td>
</tr>
<tr>
<td>M2: One-hot embeddings</td>
<td>AS</td>
<td>0.981</td>
<td>1.528</td>
<td>1.010</td>
</tr>
<tr>
<td>M3: Linear embeddings</td>
<td>AS</td>
<td>0.991</td>
<td>1.442</td>
<td>1.017</td>
</tr>
<tr>
<td>M4: DAT</td>
<td>AI</td>
<td>0.985</td>
<td>1.448</td>
<td>1.012</td>
</tr>
<tr>
<td>M5: reDAT-unsup8</td>
<td>AI</td>
<td>0.969</td>
<td>1.472</td>
<td>0.996</td>
</tr>
<tr>
<td>M6: reDAT-unsup20</td>
<td>AI</td>
<td>0.980</td>
<td>1.470</td>
<td>1.006</td>
</tr>
<tr>
<td>M7: reDAT-soft</td>
<td>AI</td>
<td>0.973</td>
<td>1.409</td>
<td>0.997</td>
</tr>
</tbody>
</table>

$^1$Normalized 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.
Results

- Results on 23K hour en-X data (normalized WER)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AS/AI</th>
<th>en-US %</th>
<th>en-GB %</th>
<th>en-AU % (unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>native</td>
<td>non-native</td>
<td>avg.</td>
</tr>
<tr>
<td>M0: Data pooling</td>
<td>AI</td>
<td>1.000</td>
<td>1.472</td>
<td>1.027</td>
</tr>
<tr>
<td>M1: AIPNet-s</td>
<td>AI</td>
<td>0.997</td>
<td>1.425</td>
<td>1.023</td>
</tr>
<tr>
<td>M2: One-hot embeddings</td>
<td>AS</td>
<td>0.981</td>
<td>1.528</td>
<td>1.010</td>
</tr>
<tr>
<td>M3: Linear embeddings</td>
<td>AS</td>
<td>0.991</td>
<td>1.442</td>
<td>1.017</td>
</tr>
<tr>
<td>M4: DAT</td>
<td>AI</td>
<td>0.985</td>
<td>1.448</td>
<td>1.012</td>
</tr>
<tr>
<td>M5: reDAT-unsup8</td>
<td>AI</td>
<td>0.969</td>
<td>1.472</td>
<td>0.996</td>
</tr>
<tr>
<td>M6: reDAT-unsup20</td>
<td>AI</td>
<td>0.980</td>
<td>1.470</td>
<td>1.006</td>
</tr>
<tr>
<td>M7: reDAT-soft</td>
<td>AI</td>
<td>0.973</td>
<td>1.409</td>
<td>0.997</td>
</tr>
</tbody>
</table>

- DAT achieves competitive WERs on both native and non-native accents but up to 13% relative WER reduction on unseen accents.
Results

- Results on 23K hour en-X data (normalized WER)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AS/AI</th>
<th>en-US % native</th>
<th>en-US % non-native</th>
<th>en-US % avg.</th>
<th>en-GB % native</th>
<th>en-GB % non-native</th>
<th>en-GB % avg.</th>
<th>en-AU % (unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: Data pooling</td>
<td>AI</td>
<td>1.000</td>
<td>1.472</td>
<td>1.027</td>
<td>1.315</td>
<td>1.574</td>
<td>1.315</td>
<td>1.393</td>
</tr>
<tr>
<td>M1: AIPNet-s</td>
<td>AI</td>
<td>0.997</td>
<td>1.425</td>
<td>1.023</td>
<td>1.330</td>
<td>1.543</td>
<td>1.332</td>
<td>1.412</td>
</tr>
<tr>
<td>M2: One-hot embeddings</td>
<td>AS</td>
<td>0.981</td>
<td>1.528</td>
<td>1.010</td>
<td>1.284</td>
<td>1.540</td>
<td>1.284</td>
<td>1.574</td>
</tr>
<tr>
<td>M3: Linear embeddings</td>
<td>AS</td>
<td>0.991</td>
<td>1.442</td>
<td>1.017</td>
<td>1.284</td>
<td>1.534</td>
<td>1.282</td>
<td>1.569</td>
</tr>
<tr>
<td>M4: DAT</td>
<td>AI</td>
<td>0.985</td>
<td>1.448</td>
<td>1.012</td>
<td>1.293</td>
<td>1.567</td>
<td>1.294</td>
<td>1.373</td>
</tr>
<tr>
<td>M5: reDAT-unsup8</td>
<td>AI</td>
<td><strong>0.969</strong></td>
<td>1.472</td>
<td><strong>0.996</strong></td>
<td><strong>1.270</strong></td>
<td><strong>1.465</strong></td>
<td><strong>1.266</strong></td>
<td><strong>1.359</strong></td>
</tr>
<tr>
<td>M6: reDAT-unsup20</td>
<td>AI</td>
<td>0.980</td>
<td>1.470</td>
<td>1.006</td>
<td>1.282</td>
<td>1.492</td>
<td>1.280</td>
<td>1.361</td>
</tr>
<tr>
<td>M7: reDAT-soft</td>
<td>AI</td>
<td>0.973</td>
<td><strong>1.409</strong></td>
<td>0.997</td>
<td>1.309</td>
<td><strong>1.440</strong></td>
<td>1.307</td>
<td>1.388</td>
</tr>
</tbody>
</table>

- DAT achieves competitive WERs on both native and non-native accents but up to 13% relative WER 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.
Results

• Results on 23K hour en-X data (normalized WER)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AS/AI</th>
<th>en-US %</th>
<th>en-GB %</th>
<th>en-AU % (unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>native</td>
<td>non-native</td>
<td>avg.</td>
</tr>
<tr>
<td>M0: Data pooling</td>
<td>AI</td>
<td>1.000</td>
<td>1.472</td>
<td>1.027</td>
</tr>
<tr>
<td>M1: AIPNet-s</td>
<td>AI</td>
<td>0.997</td>
<td>1.425</td>
<td>1.023</td>
</tr>
<tr>
<td>M2: One-hot embeddings</td>
<td>AS</td>
<td>0.981</td>
<td>1.528</td>
<td>1.010</td>
</tr>
<tr>
<td>M3: Linear embeddings</td>
<td>AS</td>
<td>0.991</td>
<td>1.442</td>
<td>1.017</td>
</tr>
<tr>
<td>M4: DAT</td>
<td>AI</td>
<td>0.985</td>
<td>1.448</td>
<td>1.012</td>
</tr>
<tr>
<td>M5: reDAT-unsup8</td>
<td>AI</td>
<td><strong>0.969</strong></td>
<td>1.472</td>
<td><strong>0.996</strong></td>
</tr>
<tr>
<td>M6: reDAT-unsup20</td>
<td>AI</td>
<td>0.980</td>
<td>1.470</td>
<td>1.006</td>
</tr>
<tr>
<td>M7: reDAT-soft</td>
<td>AI</td>
<td>0.973</td>
<td><strong>1.409</strong></td>
<td>0.997</td>
</tr>
</tbody>
</table>

• DAT achieves competitive WERs on both native and non-native accents but up to 13% relative WER 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.
• On non-native accents, reDAT with soft labels achieves significant improvements over DAT by 3% on en-US and 8% on en-GB, and over the best AI and AS baselines by 1% on en-US and 6% on en-GB.
Results

- Results on 23K hour en-X data (normalized WER)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AS/AI</th>
<th>en-US %</th>
<th>en-GB %</th>
<th>en-AU %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>native</td>
<td>non-native</td>
<td>avg.</td>
</tr>
<tr>
<td>M0: Data pooling</td>
<td>AI</td>
<td>1.000</td>
<td>1.472</td>
<td>1.027</td>
</tr>
<tr>
<td>M1: AIPNet-s</td>
<td>AI</td>
<td>0.997</td>
<td>1.425</td>
<td>1.023</td>
</tr>
<tr>
<td>M2: One-hot embeddings</td>
<td>AS</td>
<td>0.981</td>
<td>1.528</td>
<td>1.010</td>
</tr>
<tr>
<td>M3: Linear embeddings</td>
<td>AS</td>
<td>0.991</td>
<td>1.442</td>
<td>1.017</td>
</tr>
<tr>
<td>M4: DAT</td>
<td>AI</td>
<td>0.985</td>
<td>1.448</td>
<td>1.012</td>
</tr>
<tr>
<td>M5: reDAT-unsup8</td>
<td>AI</td>
<td>0.969</td>
<td>1.472</td>
<td>0.996</td>
</tr>
<tr>
<td>M6: reDAT-unsup20</td>
<td>AI</td>
<td>0.980</td>
<td>1.470</td>
<td>1.006</td>
</tr>
<tr>
<td>M7: reDAT-soft</td>
<td>AI</td>
<td>0.973</td>
<td>1.409</td>
<td>0.997</td>
</tr>
</tbody>
</table>

- DAT achieves competitive WERs on both native and non-native accents but up to 13% WER 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.
- On non-native accents, reDAT with soft labels achieves significant improvements over DAT by 3% on en-US and 8% on en-GB, and over the best AI and AS baselines by 1% on en-US and 6% on en-GB.
- In conclusion, reDAT yields significant improvements over strong baselines on non-native and unseen accents without sacrifice of native accents performance.
Conclusions

• We propose a feasible solution to mitigate accent mismatch problems for end-to-end RNN-T ASR using DAT.

• We demonstrate that DAT can achieve competitive WERs over accent-specific baselines on both native and non-native English accents, but with significantly better WER 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 Jensen-Shannon divergence between domain distributions.

• We further proposed a novel method reDAT, based on unsupervised relabeling of the training data, and obtain substantial gains over DAT on non-native accents.
Thank you!