REDAT: ACCENT-INVARIANT 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 †Amazon Alexa

[Paper No. 2760]

OVERVIEW

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.

Our work:
We aim to advance accent-invariant modeling with RNN-T based on the domain adversarial training (DAT). We propose **re-DAT**, a novel technique based on **DAT**, which **re**labels data using either unsupervised clustering or soft labels.

CONTRIBUTIONS

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

(2) We also prove that performing the gradient reversal in DAT is equivalent to minimizing the Jensen-Shannon divergence between different output distributions.

(3) We introduce reDAT, a novel technique based on DAT, which refines accent classes with either unsupervised clustering or soft labels. It yields significant improvements over strong baselines.

THEORETICAL GUARANTEE OF ACCENT-INVARIANCE FOR DAT

DAT Framework

Our DAT framework (illustrated in right part of Figure 1), consists of an accent invariant feature generator G, English accent classifier C, and RNN-T model R.
During training, weight is updated by the following gradient descent rules,

$$\theta_G \leftarrow \theta_G - \alpha\left(\frac{\partial \mathcal{L}_R}{\partial \theta_G} - \lambda \frac{\partial \mathcal{L}_C}{\partial \theta_G}\right),$$

$$\theta_C \leftarrow \theta_C - \alpha \frac{\partial \mathcal{L}_C}{\theta_C},$$

$$\theta_R \leftarrow \theta_R - \alpha \frac{\partial \mathcal{L}_R}{\theta_R}.$$

Claims: Performing gradient reversal is equivalent to minimizing Jensen-Shannon divergence (JSD) among multiple domain distributions.
The optimization rule for the accent classifier C:

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

The optimization rule for the generator G:

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

We have the following constraints for C on the probabilities $C_i(z)$ of input utterances to different accents:

$$\sum_i^N C_i(z) = 1, \qquad 0 < C_i(z) < 1.$$

Thus, we reduce the optimization of C to a linear programming problem, where C^* is the objective. It has the only solution

$$C_i^*(z) = P_{G_i} / \sum_i^N P_{G_i}.$$

Hence, for the optimization of G, by taking the optimal $C*$ into the previous expression, it can be deduced to

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

where it's equivalent to the JSD between the distributions of all accents:

$$G^* = \arg\min_{G}\left(-N\log N + JSD\left(P_{G_1}, P_{G_2}, \ldots, P_{G_N}\right)\right).$$

The global minimum is achieved if and only if P_{G_1}=P_{G_2}=...=P_{G_N}, which indicates that the embeddings z are accent-invariant. Please refer to our paper for more mathematical details.

REDAT: DAT WITH RELABELING

Motivation: From the theoretical guarantees of DAT, we could get more invariant training results by pre-defining more detailed acoustic information, such as a refined accent label for utterances.
reDAT: The DAT approach with relabeling of domain classes either by unsupervised clustering or with soft labels.

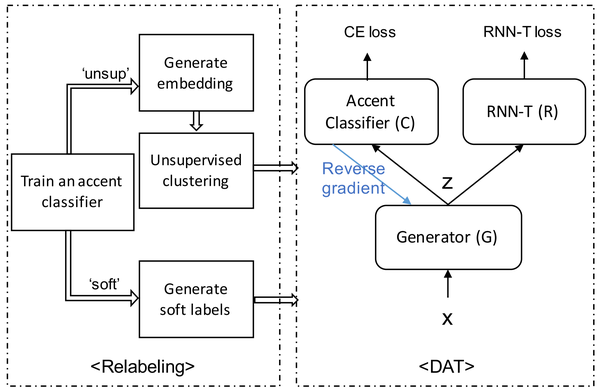

Figure 1: reDAT framework by relabeling with either unsupervised clustering ('unsup') or soft labels ('soft').

Relabeling with Unsupervised Clustering
<Step 1>: Train an accent classifier.
<Step 2>: Generate accent embeddings.
<Step 3>: Perform unsupervised clustering on accent embeddings.
<Step 4>: Perform DAT on new labels.

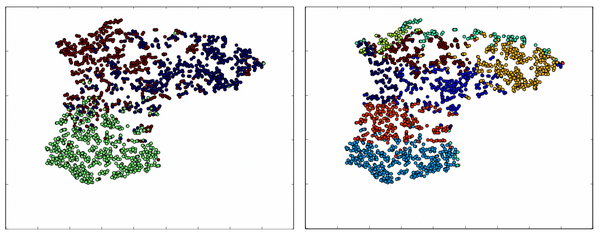

Figure 2: t-SNE visualization of unsupervised clustering on generated embeddings. Left: Target accent distribution (en-US, en-GB, en-IN). Right: K-means clustering distribution (8 classes).

Relabeling with Soft Labels
<Step 1>: Train an accent classifier.
<Step 2>: Generate soft labels
<Step 3>: Perform DAT on new labels.

EXPERIMENTS

Data set: 23K-hour en-X data: 13K hours of en-US data, 6K hours of en-GB data, and 4K hours of en-IN. en-AU is set as unseen test set.
Acoustic features: 64-dimensional log-Mel features, computed over 25ms windows with 10ms hop length. Feature vector stacked 2 frames to the left.
Model: RNN-Transducer. Encoder has 5 LSTM layers with 1024 units, prediction network has 2 LSTM layers with 1024 units.
Normalized WER: 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.

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