Domain Adversarial Training for Accented Speech Recognition

Sining Sun [1-3] , Ching-Feng Yeh [2] , Mei-Yuh Hwang [2] ,
Mari Ostendorf [3] , Lei Xie [1]

Northwestern Polytechnical University [1]
Mobvoi AI Lab, Seattle, USA [2]
University of Washington, Seattle , USA [3]

Outline

- Introduction
- Domain Adaptation
- Domain Adversarial Training (DAT)
- DAT for ASR
- Experimental Results
- Conclusion

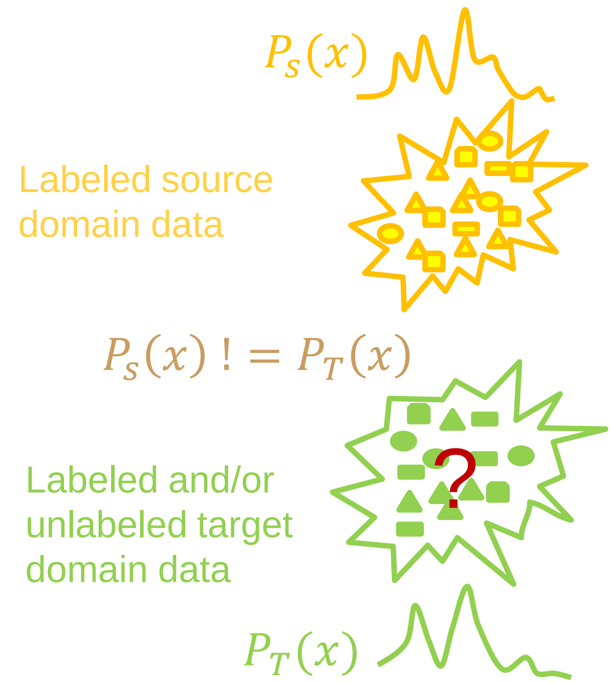
Introduction

- Challenges in ASR
  - Noise, reverberation, accents……
  - Mismatch between training and test data
  - Lack of supervised training data
- Our work
  - Improve ASR performance for accented speech, using unsupervised domain adaptation
  - Learn accent-invariant features using DAT
  - Explore how semi-supervised learning can influence the performance of DAT



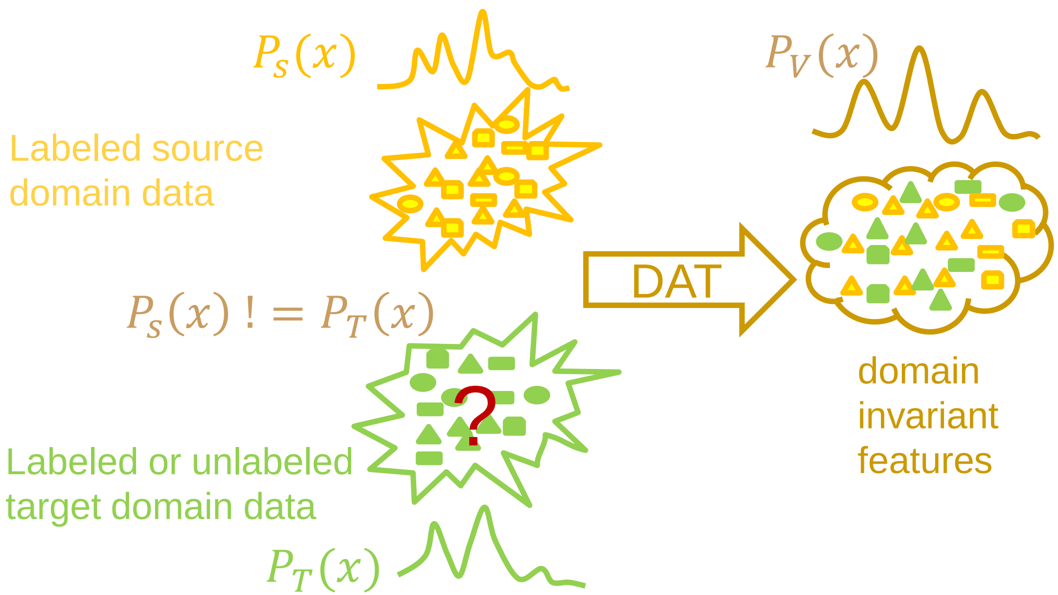
Domain Adaptation

- Domain adaptation
  - Training data
    - Labeled source domain data
    - Labeled or unlabeled target domain data
  - Test data
    - Data with the distribution of the target domain
  - Task
    - Improve performance on the test set using limited target domain data



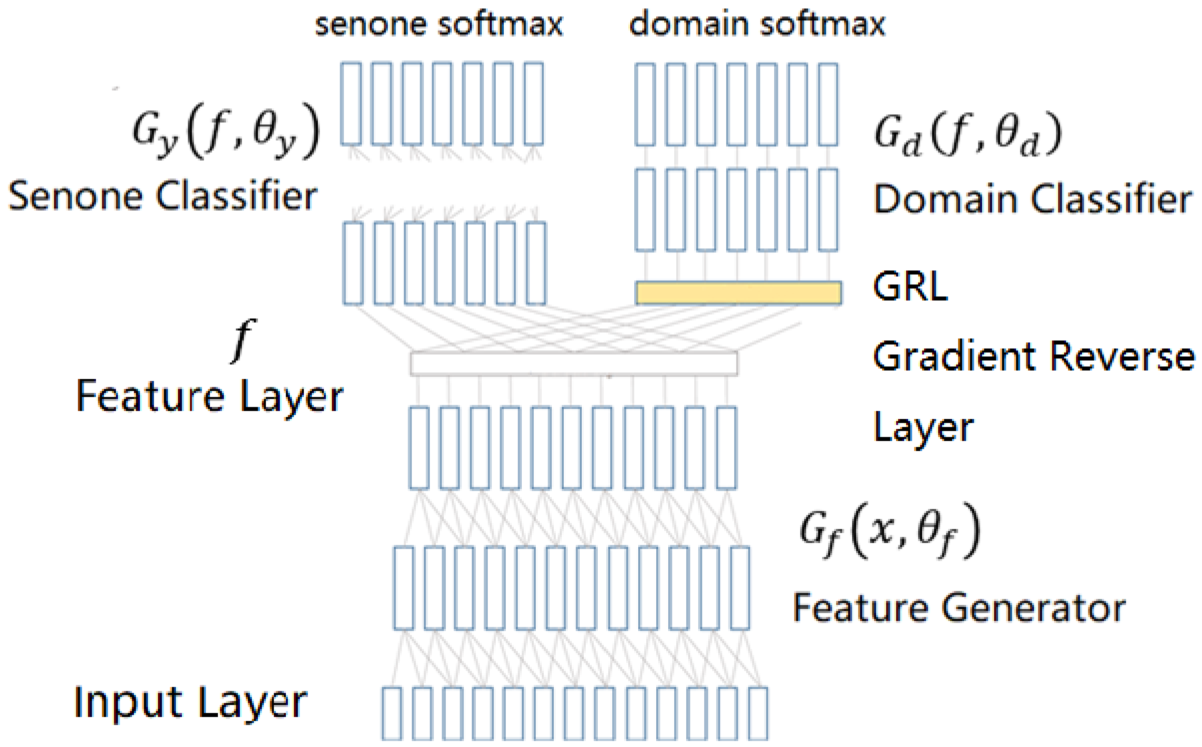
Domain Adversarial Training

- Given labeled or unlabeled target domain data
  - DAT tries to learn features that are
    - Domain-invariant
    - Classification-discriminative



DAT for Speech Recognition

- Gradient reverse layer (GRL) based adversarial training



  - GRL: multiply a constant **negative** factor ($-\lambda$) to gradients generated by $G_d(f, \theta_d)$

DAT for Speech Recognition

- Training
  - Loss function

Indicator for labeled or not

Domain cross entropy

$$E(\theta_f,\theta_y,\theta_d)=\frac{1}{N}\sum_{i=1}^{N}(I_d(i)L_y^i(\theta_f,\theta_y)-\lambda I_{vad}(i)L_d^i(\theta_f,\theta_d))$$

Senone cross entropy

Indicator for speech or not

  - Optimization

$$\theta_y^*,\theta_f^* = \min_{\theta_y,\theta_f} E(\theta_y,\theta_f,\theta_d)$$

$$\theta_d^* = \max_{\theta_d} E(\theta_y,\theta_f,\theta_d)$$

$$\Rightarrow$$

$$\theta_f \leftarrow \theta_f - \alpha\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial L_y^i}{\partial\theta_f}I_d(i)-\lambda\frac{\partial L_d^i}{\partial\theta_f}I_{vad}(i)\right)$$

$$\theta_y \leftarrow \theta_y - \alpha\frac{1}{N}\sum_{i=1}^{N}\frac{\partial L_y^i}{\partial\theta_y}I_d(i)$$

$$\theta_d \leftarrow \theta_d - \alpha\frac{1}{N}\sum_{i=1}^{N}\lambda\frac{\partial L_d^i}{\partial\theta_d}I_{vad}(i)$$

Experiment Set-up

- Dataset
  - Source domain training data
    - 360 hours standard accent Mandarin training data with transcriptions (Std)
  - Target domain training data
    - Transcribed accented Mandarin speech from:

      HaiNan (HN), SiChuan (SC), GuangDong (GD), JiangXi (JX), JiangSu (JS) and FuJian (FJ)
    - 100 hours per accent
  - Test and validation data
    - 5 hours per-accent
    - 5 hours Std data



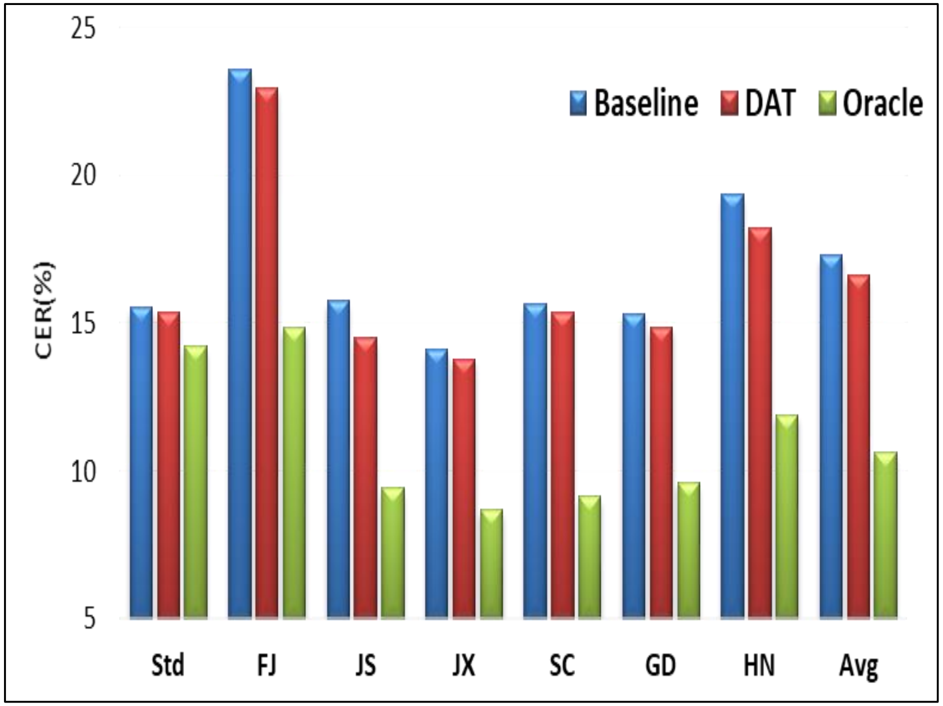
Experiment Set-up

- Acoustic feature
  - 23-dimensional filterbanks with 3-dimensional pitch
- Acoustic model
  - TDNN with LF-MMI
  - 7 layers and each layer has 625 hidden units with ReLU
  - 5998 output units
  - Trained by Kaldi
- Language model
  - 3-gram language model trained with all the text in the training set



Multi-Accent System Results

- Accent-invariant feature extraction across all accents using unsupervised DAT



Baseline:
Trained using 360 hours Std data
DAT:
Trained using 360hours Std data
\+ 600 hours accented data
without transcripts
\+ DAT
Oracle:
Trained using 360hours Std data
\+ 600 hours accented data
with human transcripts

- Using unsupervised DAT improves the ASR performance on accented test data

Per-Accent Experiments

- Three accents selected: FJ, SC, HN
- A different baseline system for each of the following conditions on 100 hours accented speech data
- Compare DAT vs MTL for different transcription cases

HN

SC

FJ

No Transcripts

Baseline: 360hrs Std data with human transcripts
DAT/MTL: 360hrs Std + 100hrs accented data *without transcripts*

ASR Transcripts

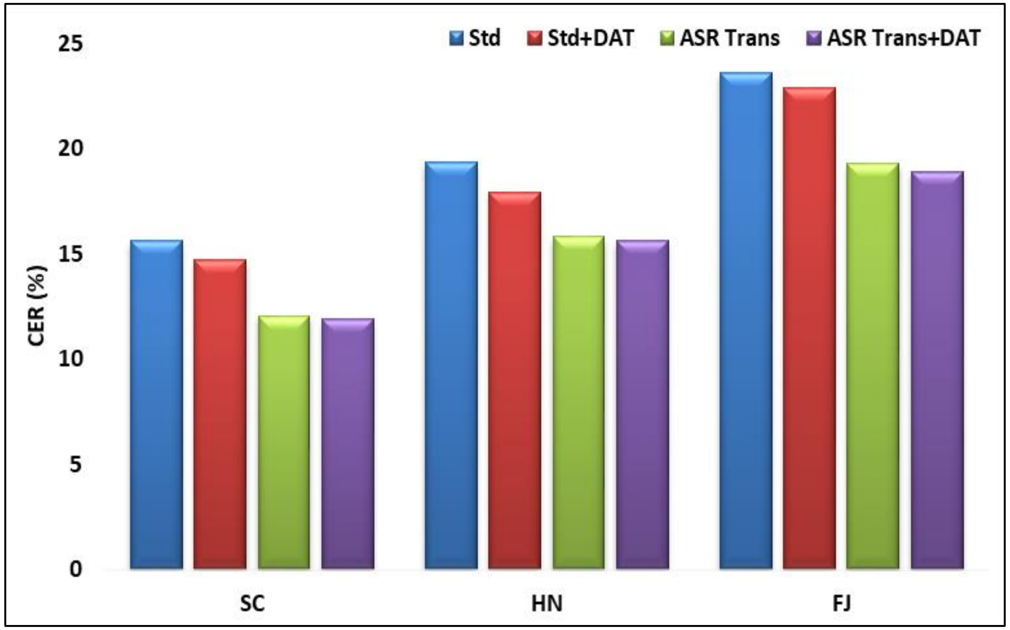
Baseline/DAT/MTL:
360hrs Std data with human transcripts
+ 100hrs accented data with *ASR transcripts*

Human Transcripts

Baseline/DAT/MTL:
360hrs Std data with human transcripts
+ 100hrs accented data with *human transcripts*

Per-Accent System Results

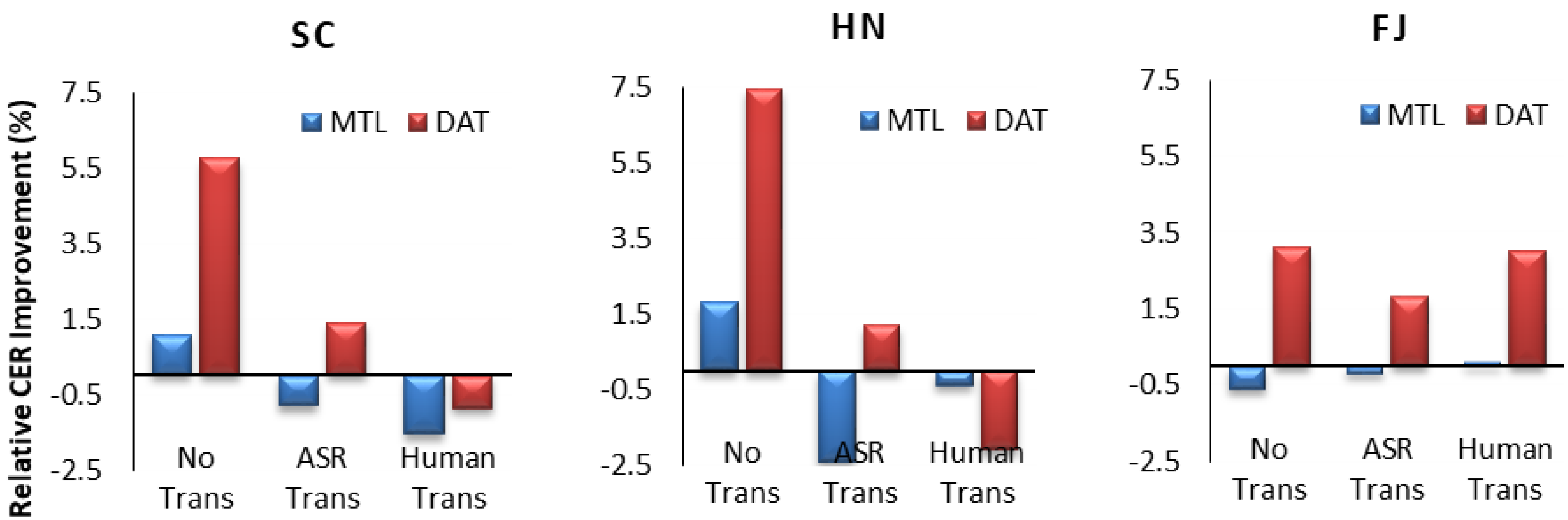
CER of different systems



- DAT alone always helps
- ASR transcripts can reduce CER further
- With ASR transcripts, DAT helps, but the contribution shrinks

DAT vs MTL

Relative CER improvement of accent-specific DAT



- When no transcript or ASR transcripts were available, DAT always helps
- DAT is always better than MTL

Conclusion

- Conclusion
  - Integrated DAT into TDNN AM training for accented speech recognition
  - 7.4% relative CER reduction using unsupervised DAT
  - Explored how automatic transcripts influence DAT performance
  - 20% relative CER reduction when combining DAT and ASR transcripts
- Future work
  - Compare DAT with other emerging deep domain adaption methods
  - Extend DAT to far-field scenario

Thank you!