Unsupervised Domain Adaptation for Robust Speech Recognition via Variational Autoencoder-Based Data Augmentation

Wei-Ning Hsu, Yu Zhang, James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139

Summary

- Proposed a novel data-driven non-heuristic data augmentation method for unsupervised domain adaptation, which requires zero in-domain labeled data.
- Achieved up to 35% and 40% absolute word error rate reduction in mismatched domains on CHiME-4 and Aurora-4 respectively.

Robust Automatic Speech Recognition

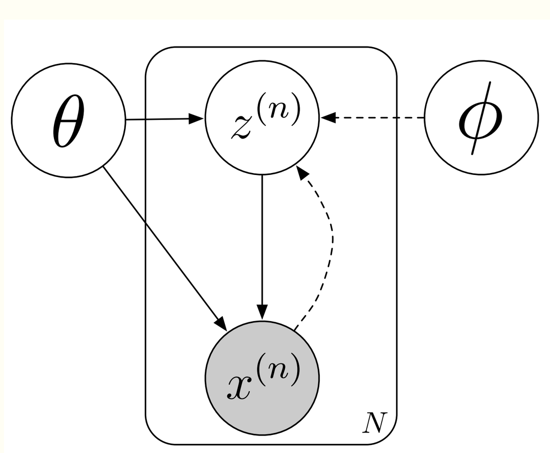
An ASR system often degrades significantly when testing on a domain mismatched from the training data. Here are a few ways to achieve robustness:

- use domain-invariant acoustic features.
- enhance speech (convert out-of-domain data to in-domain data).
- train an ASR system with as much, and as diverse a dataset as possible.

⇒ **use more data to utilize the full capacity of neural network models.**

Unsupervised Learning of Latent Factors with VAEs

Variational Autoencoders (VAEs)



Consider a speech dataset $\mathcal{D} = \{x^{(n)}\}_{n=1}^{N}$ of N i.i.d. speech segments. Each x is assumed to be generated by:

1. draw a **latent variable** z from $p_\theta(z) = \mathcal{N}(z|\mathbf{0}, \mathbf{I})$
2. draw an **observed variable** x from $p_\theta(x|z) = \mathcal{N}(x|f_{\mu_x}(z), \exp(f_{\log\sigma_x}(z)))$

A variational inference model $q_\phi(z|x)$ is introduced to approximate the intractable true posterior $p_\theta(z|x)$

- $q_\phi(z|x) = \mathcal{N}(x|g_{\mu_z}(x), \exp(g_{\log\sigma_z}(x)))$

Objective Function: **Variational Lower Bound**

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) \quad (1)$$

Properties of Latent Variables:

- The prior $p_\theta(z)$ is a factorial distribution.
- VAEs are encouraged to encode independent physical attributes (e.g. speaker identity, phonemes) into orthogonal subspaces.

Latent Attribute Representation

It is suggested and empirically verified in the previous work [Hsu et. al., 2017]:

1. Conditional prior of z on some attribute a being r (e.g. *phoneme* being */aa/*):

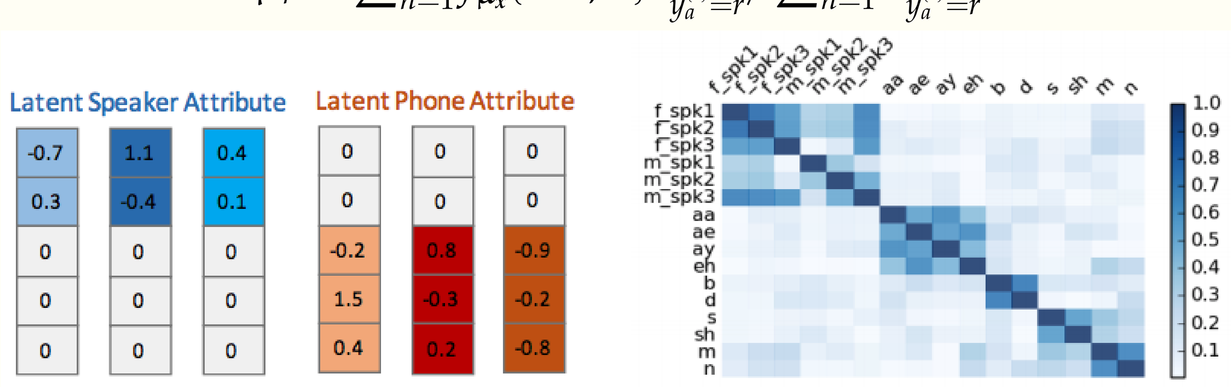
$$p_\theta(z|y_a = r) = \mathcal{N}(z; \mu_r, \Sigma_r) \quad (2)$$

2. $\mu_{r_i} \perp \mu_{r_j}$ if r_i and r_j are independent attributes that affect the realization of speech.

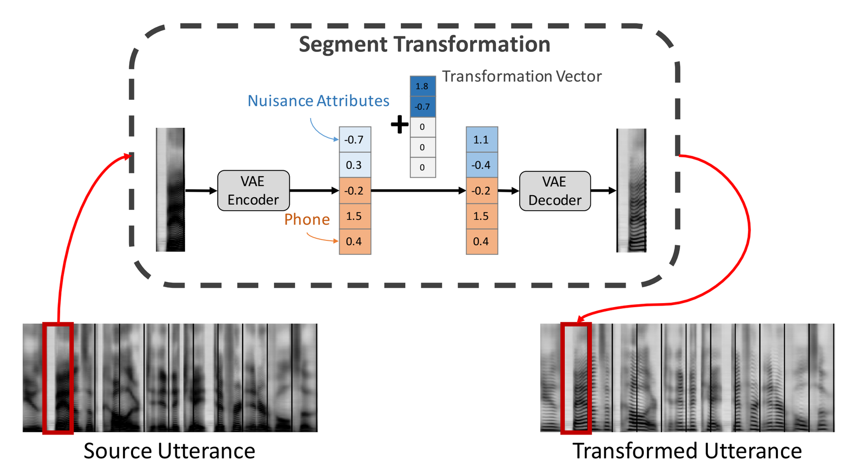
⇒ μ_r is defined as **latent attribute representation** for r.

Estimating Latent Attribute Representations

$$\mu_r \approx \sum_{n=1}^{N} f_{\mu_z}(x^{(n)}; \theta)\mathbf{1}_{y_a^{(i)}=r} / \sum_{n=1}^{N} \mathbf{1}_{y_a^{(i)}=r}$$



Transforming an Utterance



- given the orthogonality property, one can modify some attributes of a speech utterance without changing other independent attributes.

Nuisance Attributes and Data Augmentation

Nuisance Attributes: factors that affect the surface form of a speech utterance but not the linguistic content, such as *speaker identity*, *channel*, *background noise*.

- nuisance attributes are independent from linguistic content.

⇒ **Given a labeled utterance, (1) encode, (2) modify the latent subspace that models these attributes, and (3) decode, to generate augmented labeled data**

Estimating Latent Nuisance Representations:

- nuisance attributes are generally consistent within an utterance.
- same labels for these attributes for all the segments within an utterance.
- suppose $\{x_{utt_j}^{(n)}\}_{n=1}^{N_j}$ be the set of segments from an utterance utt_j, we then have:

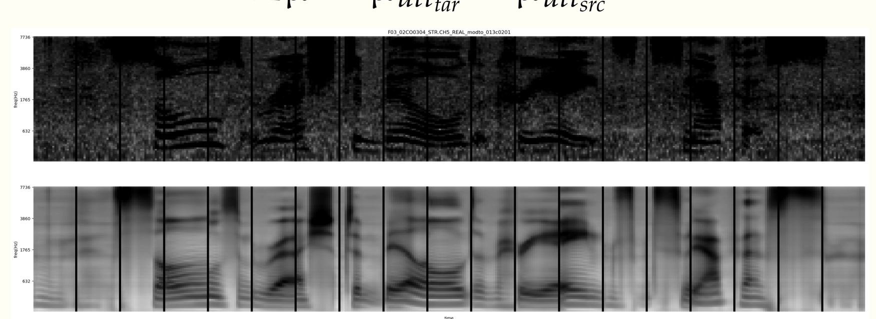
$$\mu_{utt_j} = \sum_{n=1}^{N_j} f_{\mu_z}(x_{utt_j}^{(n)}; \theta)/N_j \quad (3)$$

Type I: Nuisance Attribute Replacement (Repl.)

Replace the nuisance attribute of one utterance with that of another utterance.

Transformation vector:

$$\Delta\mu = \mu_{utt_{tar}} - \mu_{utt_{src}} \quad (4)$$

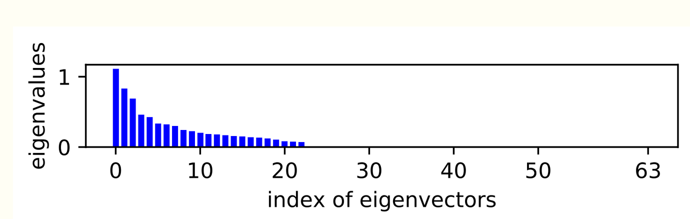


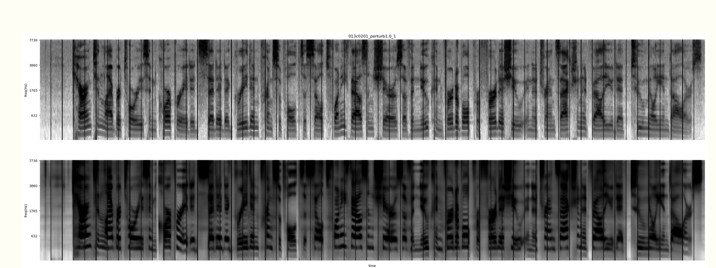
Type II: Latent Nuisance Subspace Perturbation (Pert.)

Discover the latent nuisance subspace and perturb that subspace

1. determine latent nuisance subspace with principle component analysis (PCA)
   - given a dataset of M utterance, we can compute $\{\mu_{utt_i}\}_{i=1}^{M}$, on which we apply PCA.
   - obtain a list of eigenvectors $\{e_d\}_{d=1}^{D}$ with associated eigenvalues $\{\sigma_d^2\}_{d=1}^{D}$.
2. sample a transformation vector for soft latent nuisance subspace perturbation:

$$\Delta\mu = \gamma \sum_{d=1}^{D} \phi_d \sigma_d e_d, \quad \phi_d \sim \mathcal{N}(0, 1) \quad (5)$$





Experiment Setup

Datasets

- **CHiME-4:** the training set consists of 1600 real noisy utterances and 7138 WSJ0 SI-84 clean utterances
- **Aurora-4:** multi-condition speech dataset, 2 microphone types, 6 noise types, 4620 WSJ-0 based utterances.

VAE Model

- **Input:** x is a segment of 20 frames, represented as mel-scale filter bank coefficient (FBank)
- **Encoder/Decoder:** two layer LSTM with 512 hidden units. Adam optimizer.
- **Training Set:** all conditions (clean + noisy)

ASR Acoustic Model

- transcripts available for only clean set. augmentation is based on clean data.
- three layer LSTM with 1024 cells and 512-node projection (CHiME-4) / six layer fully-connected with 1024 hidden units (Aurora-4)

CHiME-4 Results

Setting			WER (%)		WER (%) by Environment			
Exp. Index	Aug. Method	Fold	Clean	Noisy	BUS	CAF	PED	STR
1	Orig.	1	19.04	87.80	96.16	92.35	78.46	84.24
2	Repl. Clean	1	**20.03**	67.12	71.99	76.84	55.32	64.33
	Repl. Noisy	1	26.31	**57.66**	62.12	69.25	46.89	52.38
3	Pert., $\gamma = 1.0$	1	20.01	**53.06**	55.66	66.12	41.94	48.50
	Uni-Pert., $\gamma = 1.0$	1	**19.70**	65.07	69.27	75.28	53.65	62.06
	Rev-Pert., $\gamma = 1.0$	1	19.75	87.98	95.13	90.58	76.71	89.50
4	Pert., $\gamma = 0.5$	1	**19.55**	65.61	67.87	77.37	54.54	62.66
	Pert., $\gamma = 1.0$	1	20.01	**53.06**	55.66	66.12	41.94	48.50
	Pert., $\gamma = 1.5$	1	19.99	53.59	57.09	64.91	42.23	50.11
5	Orig. + Repl. Noisy	2	19.88	55.72	60.72	66.46	45.08	50.63
	Repl. Noisy	2	25.26	55.59	59.24	67.85	44.65	50.63
	Pert., $\gamma = 1.0$	2	**19.82**	**52.49**	55.52	65.04	41.17	48.24

Table: CHiME-4 development set word error rate of acoustic models trained on different augmented sets.

We showed the following results:

- correctness of soft latent nuisance subspace perturbation
- effectiveness of both replacement and perturbation, and superiority of the latter.
- benefit of generating more augmented data

Aurora-4 Results

Setting			WER (%)	WER (%) by Condition			
Exp. Index	Aug. Method/Baselines	Fold	Avg.	Cln	Noisy	Chan	N+Ch
0	Clean-DNN-HMM	-	36.22	3.36	29.74	21.02	50.73
	DDA-DNN-HMM	-	22.53	3.24	14.52	17.82	34.55
	DNN-PP	-	18.7	5.1	12.0	10.5	29.0
1	Orig.	1	53.98	3.38	50.56	42.67	67.70
2	Repl. Noisy	1	22.53	4.80	16.31	14.72	32.99
3	Pert., $\gamma = 2.0$	1	20.68	4.45	14.33	14.74	30.72
4	Pert., $\gamma = 2.0$	16	**18.76**	**4.04**	**12.84**	**13.54**	**28.01**

Table: Aurora-4 test_eval92 set word error rate of acoustic models trained on different augmented sets.

- Outperformed a state-of-the-art domain adversarial training-based method (DDA).
- Matched the performance of an enhancement-based method (DNN-PP), which however requires parallel data.

wnhsu@mit.edu