Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition

Wei-Ning Hsu, James Glass

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

Summary

- Proposed an unsupervised learning framework for extracting domain invariant ASR features using factorized hierarchical variational autoencoders (FHVAEs).
- Achieved up to 41% and 27% absolute word error rate reductions in mismatched domains on CHiME-4 and Aurora-4.

Robust Automatic Speech Recognition

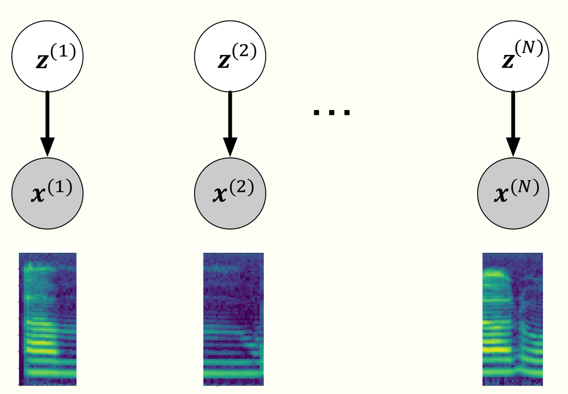
An ASR system often degrades significantly when testing on a domain mismatched from the training data. Here are a few ways to achieve robustness:

1. multi-condition training.
2. transform training or testing data. (corrupting training data or enhancing testing data)
3. use domain-invariant acoustic features.

⇒ **1. often requires labeled data in all domains, and 2. often requires parallel data between domains. We investigate 3. that has no such constraints.**

Unsupervised Learning of Domain Invariant Features

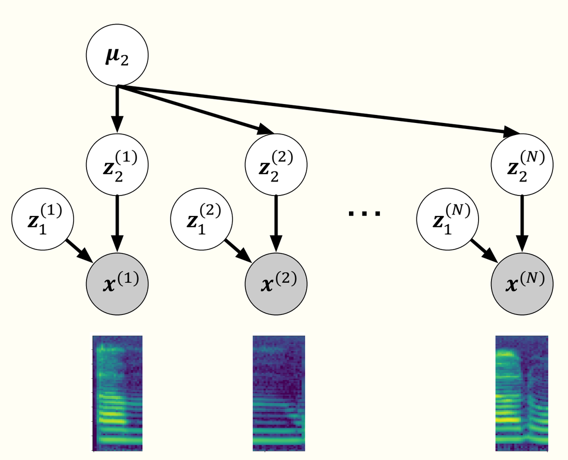
Background: Variational Autoencoders (VAEs)



- Describe data generation using a directed graphical model with a latent variable z.
- Define a *decoder neural network* to parameterize $p(x|z)$.
- Define an *encoder neural network* to parametrize $q(z|x)$, an amortized approximation of the intractable $p(z|x)$.
- The encoder/decoder networks are trained jointly to maximize a lower bound of the marginal likelihood $p(x)$, called the *variational lower bound* $\mathcal{L}(x; p, q)$.

⇒ Learns a representation encoding *all* generating factors, not domain invariant

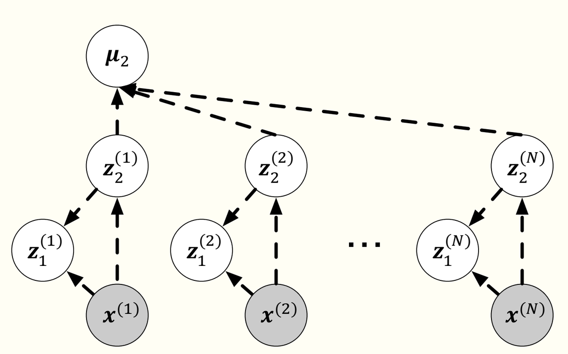
Factorized Hierarchical Variational Autoencoders (FHVAEs)



- Describe sequential data generation using a directed hierarchical graphical model with latent variables z_1, z_2, and μ_2.

$$\mu_2 \sim \mathcal{N}(0, I) \quad (1)$$
$$z_1 \sim \mathcal{N}(0, I) \quad (2)$$
$$z_2 \sim \mathcal{N}(\mu_2, \sigma^2 I) \quad (3)$$
$$x \sim \mathcal{N}(\mathrm{dec}_{\mu}(z_1, z_2), \mathrm{dec}_{\sigma^2}(z_1, z_2)) \quad (4)$$

- μ_2 is **shared** for segments from the same sequence.
- z_2 within a sequence is encouraged to be close to each other. ⇒ encode static generating factors.
- z_1 captures residual time-varying generating factors.



- Use encoder networks for variational inference

$$z_2|x \sim \mathcal{N}(z_2\text{-}\mathrm{enc}_{\mu}(x), z_2\text{-}\mathrm{enc}_{\sigma^2}(x)) \quad (5)$$
$$z_1|x, z_2 \sim \mathcal{N}(z_1\text{-}\mathrm{enc}_{\mu}(x, z_2), z_1\text{-}\mathrm{enc}_{\sigma^2}(x, z_2)) \quad (6)$$
$$\mu_2|\{z_2^{(i)}\}_{n=1}^N \sim \mathcal{N}\left(\frac{\sum_{n=1}^N z_2^{(n)}}{N + \sigma^2}, I\right) \quad (7)$$

⇒ An FHVAE learns a disentangled representation

- Domain-related factors are encoded by z_2, as such factors are static within an utterance.
- Domain-invariant phonetic factors are encoded by z_1, which are time-varying within an utterance.

Experiment Setup

We evaluate domain invariance of features by training a supervised model on one domain with different features, and testing on multiple domains.

⇒ smaller testing performance gap between domains indicates better invariance.

Datasets

- **Aurora-4:** synthesized noisy + clean
- **CHiME-4:** real noisy + clean

ASR Acoustic Model

- **Model**: three layer LSTM with 1024 cells and 512-node projection
- **Training Set:** clean
- **Objective:** frame-level cross entropy

FHVAE/VAE Model

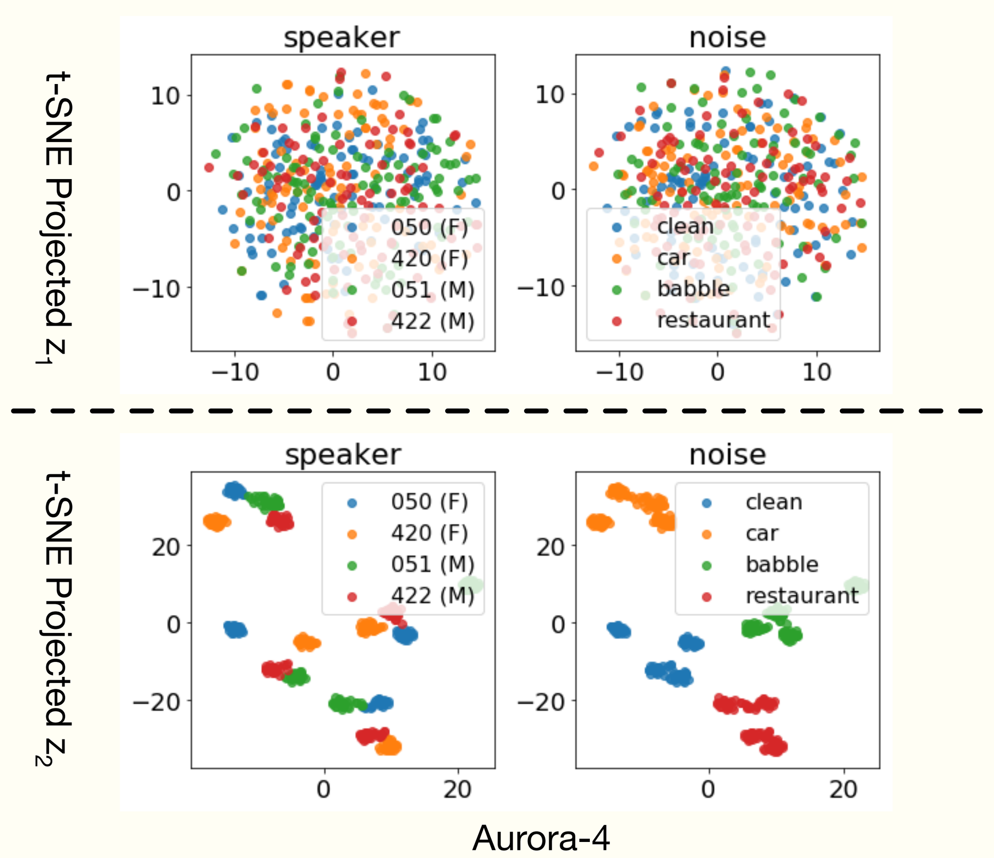
- **Input:** x is a segment of 20 frames, represented as mel-scale filter bank coefficient (FBank)
- **Encoder/Decoder:** Seq2Seq LSTM with 1/2/3 layers and 128/256/512 cells
- **Training Set:** clean + noisy
- **Objective:** Discriminative Segmental Variational Lower Bound

Discriminative Segmental Variational Lower Bound:

$$\mathcal{L}^{dis}(p, q; x^{(i,n)}) = \mathcal{L}(p, q; x^{(i,n)}) + \alpha \log \frac{p(\bar{z}_2^{(i,n)}|\bar{\mu}_2^{(i)})}{\sum_{j=1}^{M} p(\bar{z}_2^{(i,n)}|\bar{\mu}_2^{(j)})}; \quad \bar{z}, \bar{\mu} \text{ posterior means} \quad (8)$$

Qualitative Study: t-SNE visualization

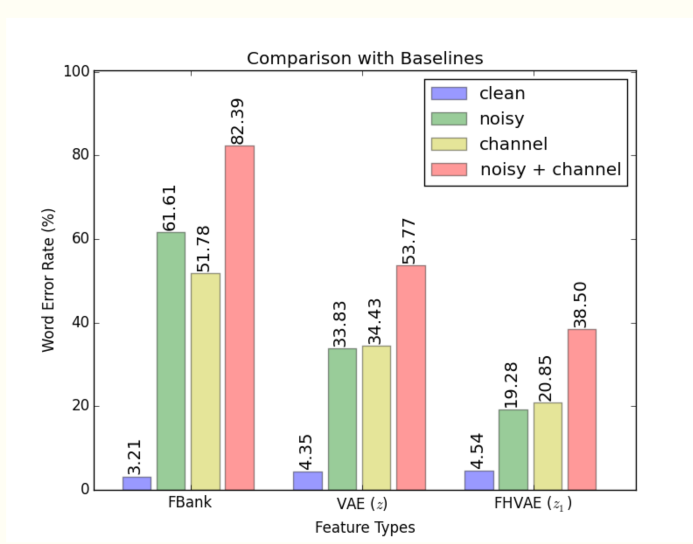
We sample segments of **4 noise types** and **4 speakers**, infer their latent variable z_1 and z_2, and use t-SNE to project z_1 and z_2 to two-dimensional spaces respectively.



- Domain-related information, such as speaker and noise type, is clearly encoded by z_2.
- Conditional distributions of projected z_1 of different domains do not seem to vary.

ASR Results Comparing with Baselines (Aurora-4)

Baseline Features: FBank / VAE latent variable z

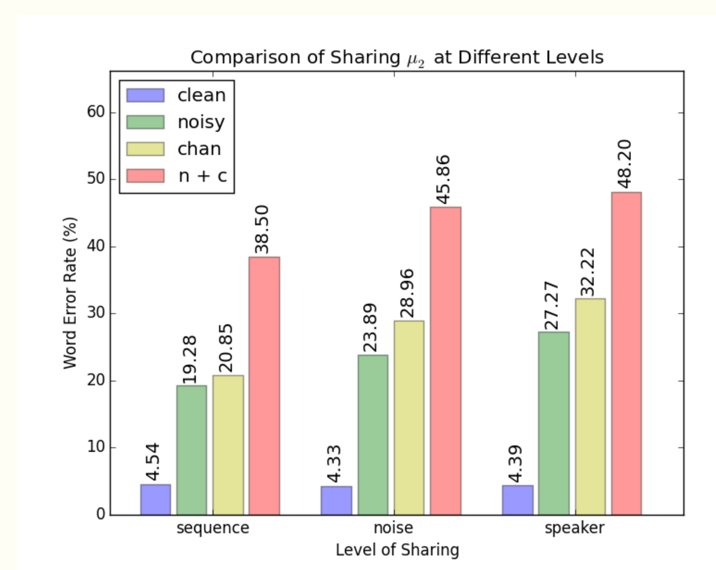


- FBank degrades significantly (49% to 79% absolute) in mismatched domains.
- Having been trained on both clean and noisy data, VAE features suffer less degradation. However such features still contain domain information.
- FHVAE features consistently outperform two baselines in all mismatched domains by a large margin, showing better domain resistance.

Shared μ_2 at Speaker Or Noise Level

Recall that μ_2 is shared at the sequence level in the original FHVAE formulation.

- With speaker or noise label available, we can share μ_2 at the speaker or noise level.

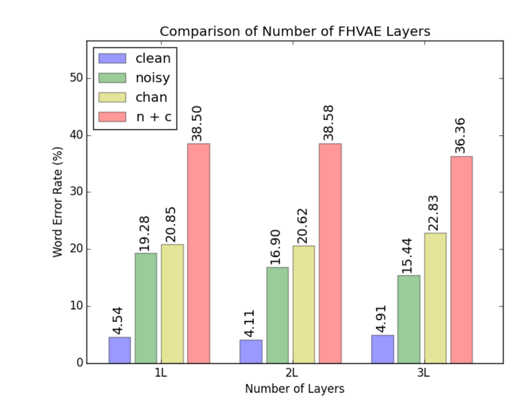


- Surprisingly, utilizing speaker/noise label in such way deteriorates the performance.
- Reasons are that when sharing μ_2 at the speaker level, noise is not a static generating factor anymore, which would then be encoded by z_1.
- This also explains sharing at the speaker level results in worse performance than sharing at the noise-type level.

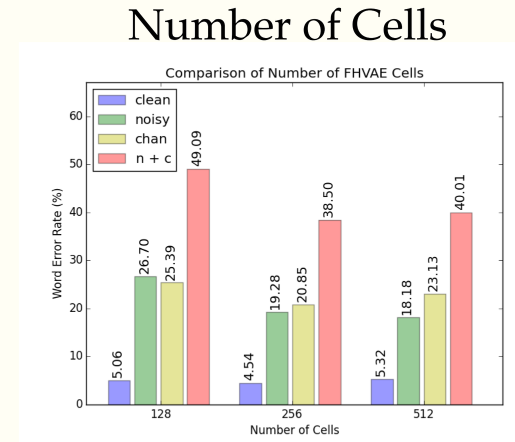
Extensive Hyper-parameter Search

We proceed with hyper-parameter search for FHVAE models:

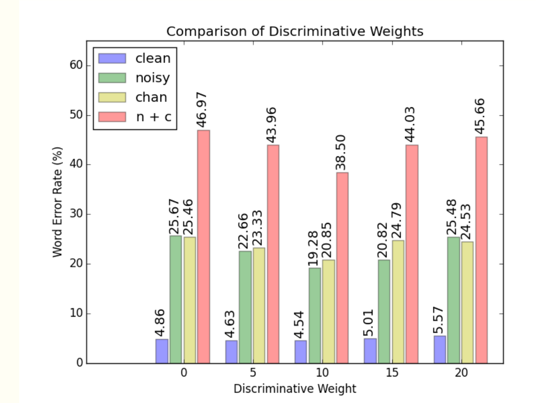
Number of Layers



Number of Cells



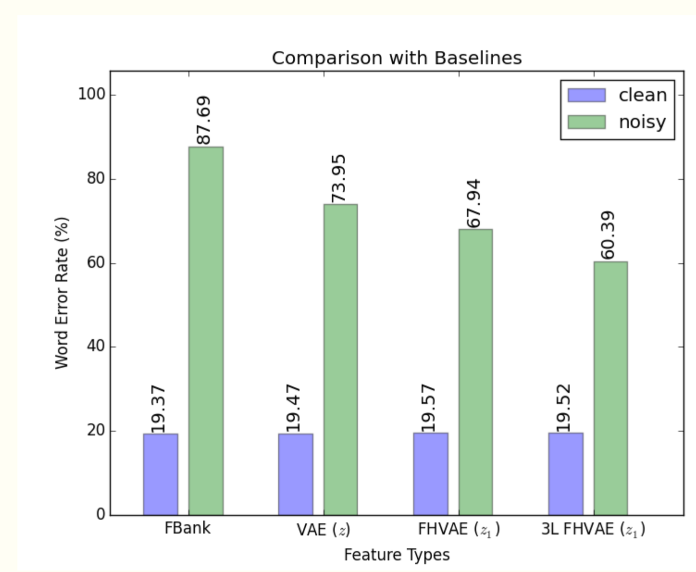
Discriminative Weights

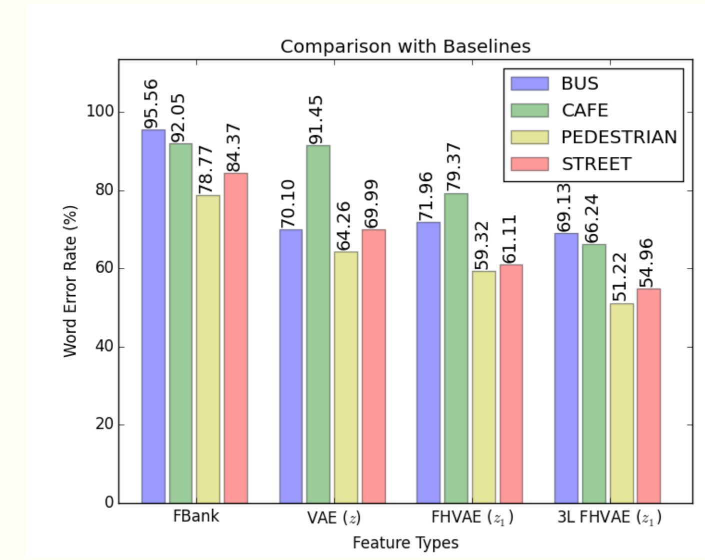


⇒ **3-Layer, 256-cells FHVAE trained with discriminative loss $\alpha = 10$ yields the best performance.**

Verifying ASR Results on CHiME-4

Baseline Features: FBank / VAE latent variable z





- FHVAE features outperform both baselines, consistent with results on Aurora-4
- Increasing number of FHVAE layers from 1 to 3 shows further improvement.

Future Work

- Investigate data augmentation-based methods using FHVAEs.
- Combining domain invariant features with adversarial training for acoustic models to further boost the robustness.

wnhsu@mit.edu