Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data

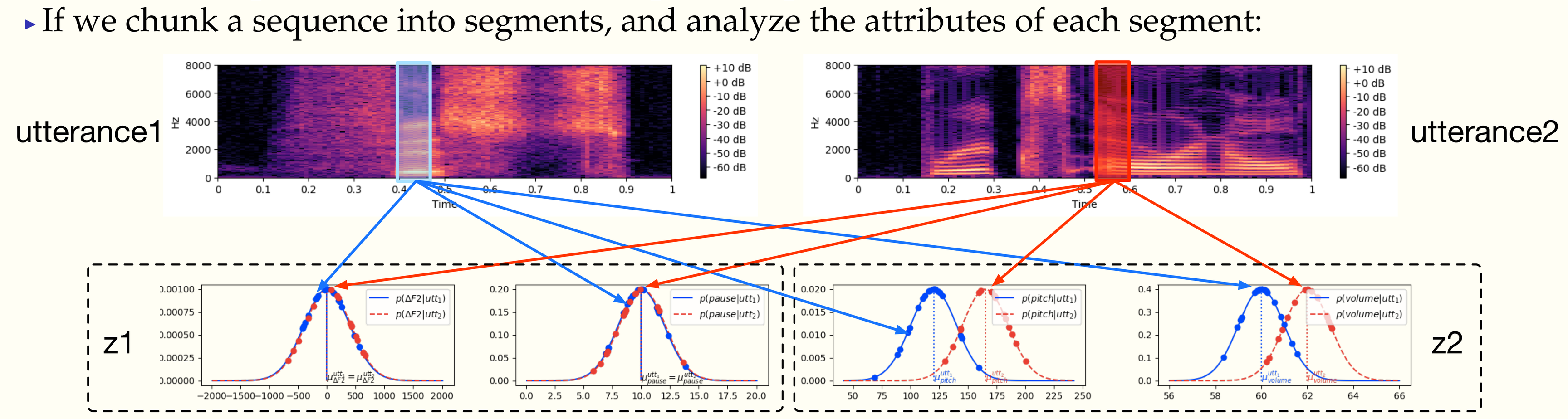
Wei-Ning Hsu, Yu Zhang, James Glass
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139

Summary

- Proposed a factorized hierarchical variational autoencoder (FHVAE) model, which learns to encode sequence-level attributes (e.g. speaker) and segment-level attributes (e.g. phoneme) into different sets of latent variables.
- Showed the capability of voice conversion and denoising without parallel data.
- Applied learned representations to speaker verification and domain invariant speech recognition tasks, which outperform i-vectors and reduce word error rate by 35%.

Motivation

- Generation of sequential data involves multiple independent factors operating at different temporal scales (channel/speaker/phoneme).
  - If we chunk a sequence into segments, and analyze the attributes of each segment:

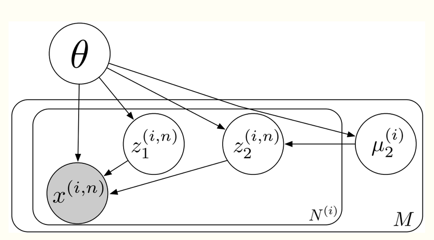


- Some attributes tend to have a smaller amount of variation within a sequence, compared to between utterances, such as **F0** and **volume** → **Sequence-Level Attributes**
- Other attributes tend to have a similar amount of variation within and between utterances, such as **phonetic content** → **Segment-Level Attributes**

⇒ We can exploit this property to factorize sequence-level and segment-level attributes

Factorized Hierarchical Variational Autoencoders

- A generative process for a sequence $\boldsymbol{X} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$:

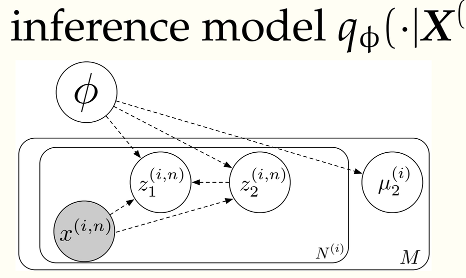


1. draw an **s-vector** μ_2 from $p_\theta(\mu_2) = \mathcal{N}(\mu_2|\mathbf{0}, \sigma^2_{\mu_2}\boldsymbol{I})$
2. draw N i.i.d. **latent sequence variables** $\boldsymbol{Z}_2 = \{\boldsymbol{z}_2^{(n)}\}_{n=1}^{N}$ from $p_\theta(\boldsymbol{z}_2|\mu_2) = \mathcal{N}(\boldsymbol{z}_2|\mu_2, \sigma^2_{z_2}\boldsymbol{I})$
3. draw N i.i.d. **latent segment variables** $\boldsymbol{Z}_1 = \{\boldsymbol{z}_1^{(n)}\}_{n=1}^{N}$ from $p_\theta(\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_1|\mathbf{0}, \sigma^2_{z_1}\boldsymbol{I})$.
4. draw N i.i.d. **observed variables** $\boldsymbol{X} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ from $p_\theta(\boldsymbol{x}|\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathcal{N}(\boldsymbol{x}|f_{\mu_x}(\boldsymbol{z}_1, \boldsymbol{z}_2), diag(f_{\sigma^2_x}(\boldsymbol{z}_1, \boldsymbol{z}_2)))$.

- Joint probability:

$$p_\theta(\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \mu_2) = p_\theta(\mu_2) \prod_{n=1}^{N} p_\theta(\boldsymbol{x}^{(n)}|\boldsymbol{z}_1^{(n)}, \boldsymbol{z}_2^{(n)}) p_\theta(\boldsymbol{z}_1^{(n)}) p_\theta(\boldsymbol{z}_2^{(n)}|\mu_2)$$

- An inference model $q_\phi(\cdot|\boldsymbol{X}^{(i)})$ for approximating $p_\theta(\cdot|\boldsymbol{X}^{(i)})$ (i is a sequence index):



  - $q_\phi(\mu_2^{(i)}) = \mathcal{N}(\mu_2^{(i)}|g_{\mu_{\mu_2}}(i), \sigma^2_{\tilde{\mu}_2}\boldsymbol{I})$
  - $q_\phi(\boldsymbol{z}_2|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_2|g_{\mu_{z_2}}(\boldsymbol{x}), diag(g_{\sigma^2_{z_2}}(\boldsymbol{x})))$
  - $q_\phi(\boldsymbol{z}_1|\boldsymbol{x}, \boldsymbol{z}_2) = \mathcal{N}(\boldsymbol{z}_1|g_{\mu_{z_1}}(\boldsymbol{x}, \boldsymbol{z}_2), diag(g_{\sigma^2_{z_1}}(\boldsymbol{x}, \boldsymbol{z}_2)))$

- Posterior probability:

$$q_\phi(\boldsymbol{Z}_1^{(i)}, \boldsymbol{Z}_2^{(i)}, \mu_2^{(i)}|\boldsymbol{X}^{(i)}) = q_\phi(\mu_2^{(i)}) \prod_{n=1}^{N^{(i)}} q_\phi(\boldsymbol{z}_1^{(i,n)}|\boldsymbol{x}^{(i,n)}, \boldsymbol{z}_2^{(i,n)}) q_\phi(\boldsymbol{z}_2^{(i,n)}|\boldsymbol{x}^{(i,n)})$$

- Objective Function: **Segment Variational Lower Bound**

$$\begin{aligned}
\mathcal{L}(\theta, \phi; \boldsymbol{x}^{(n)}) &= \mathcal{L}(\theta, \phi; \boldsymbol{x}^{(n)}|\tilde{\mu}_2) + \frac{1}{N}\log p_\theta(\tilde{\mu}_2) + const \\
\mathcal{L}(\theta, \phi; \boldsymbol{x}^{(n)}|\tilde{\mu}_2) &= \mathbb{E}_{q_\phi(\boldsymbol{z}_1^{(n)}, \boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})}\left[\log p_\theta(\boldsymbol{x}^{(n)}|\boldsymbol{z}_1^{(n)}, \boldsymbol{z}_2^{(n)})\right] \\
&\quad - \mathbb{E}_{q_\phi(\boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})}\left[D_{KL}(q_\phi(\boldsymbol{z}_1^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{z}_2^{(n)})||p_\theta(\boldsymbol{z}_1^{(n)}))\right] - D_{KL}(q_\phi(\boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})||p_\theta(\boldsymbol{z}_2^{(n)}|\tilde{\mu}_2)) \\
\tilde{\mu}_2 &= g_{\mu_{\mu_2}}(i)
\end{aligned}$$

Boosting Factorization with Discriminative Objective

- We do not want $\tilde{\mu}_2$ for different sequences to collapse to the same mode
  ⇒ FHVAE would degenerate to normal VAE in this case
- Encourage **discriminability of z_2** regarding sequences:

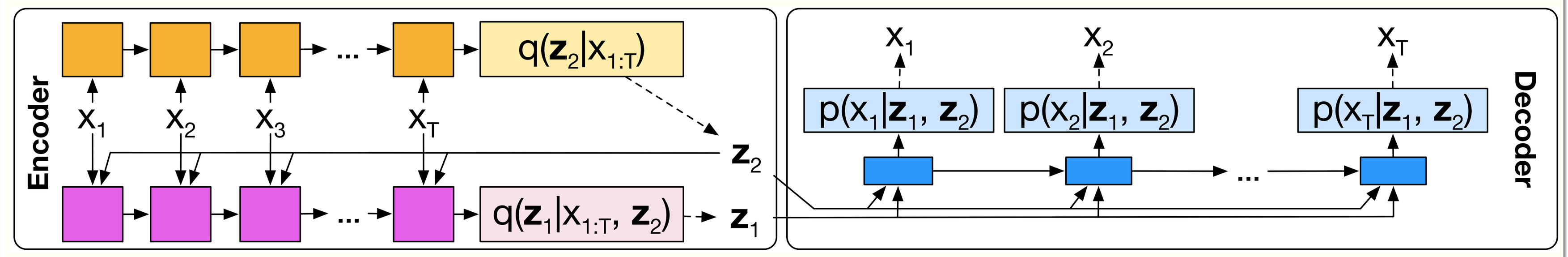
$$\log p(i|\boldsymbol{z}_2^{(i,n)}) = \log p(\boldsymbol{z}_2^{(i,n)}|i) - \log \sum_{j=1}^{M} p(\boldsymbol{z}_2^{(i,n)}|j) := \log p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\mu}_2^{(i)}) - \log \left(\sum_{j=1}^{M} p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\mu}_2^{(j)}) \right)$$

- New Objective Function: **Discriminative Segment Variational Lower Bound**

$$\mathcal{L}^{dis}(\theta, \phi; \boldsymbol{x}^{(i,n)}) = \mathcal{L}(\theta, \phi; \boldsymbol{x}^{(i,n)}) + \alpha \log p(i|\boldsymbol{z}_2^{(i,n)})$$

Segment-to-Segment FHVAE Architecture

- Each $\boldsymbol{x}$ is a **segment** (sub-sequence) of a sequence $\boldsymbol{X}$.
  - we need an *encoder* to infer z_1 and z_2 from a segment $\boldsymbol{x}$ ($g_{\mu_{z_2}}(\cdot)$, $g_{\sigma^2_{z_2}}(\cdot)$, $g_{\mu_{z_1}}(\cdot,\cdot)$, and $g_{\sigma^2_{z_1}}(\cdot,\cdot)$),
  - and a *decoder* to generate a segment $\boldsymbol{x}$ conditioned on z_1 and z_2 ($f_{\mu_x}(\cdot,\cdot)$ and $f_{\sigma^2_x}(\cdot,\cdot)$)).
- We apply a **segment-to-segment model**. Let $\boldsymbol{x} = \{x_t\}_{t=1}^{T}$ be a segment of T time steps:



Experiment Setup

Datasets

- **TIMIT:** clean speech dataset, 6300 utterances (5.4 hours), 630 speakers.
- **Aurora-4:** multi-condition speech dataset, 2 microphone types, 6 noise types, 4620 WSJ-0 based utterances (9 hours)

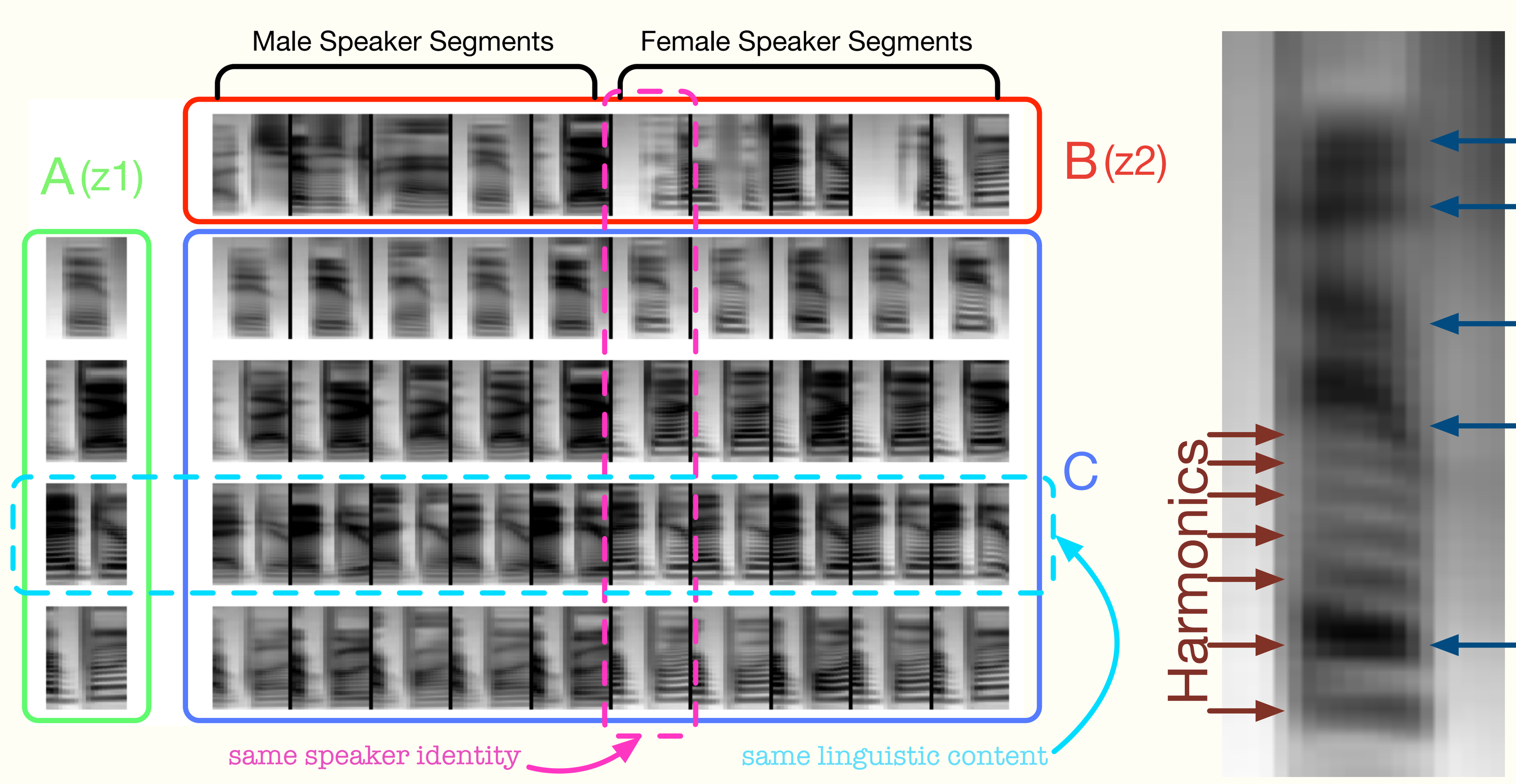
Model

- **Input:** $\boldsymbol{x}$ is a segment of 20 frames, represented as mel-scale filter bank coefficient (FBank) or log power spectrum. Feature frames are computed every 10ms
- **Encoder/Decoder:** 256 hidden units, $\sigma^2_{z_1} = \sigma^2_{\mu_2} = 1$, $\sigma^2_{z_2} = 0.25$, ADAM optimizer.

Qualitative Evaluation – Visualize Factorization

Generate a segment **C**, conditioned on the **latent segment variable of A** and the **latent sequence variable of B**

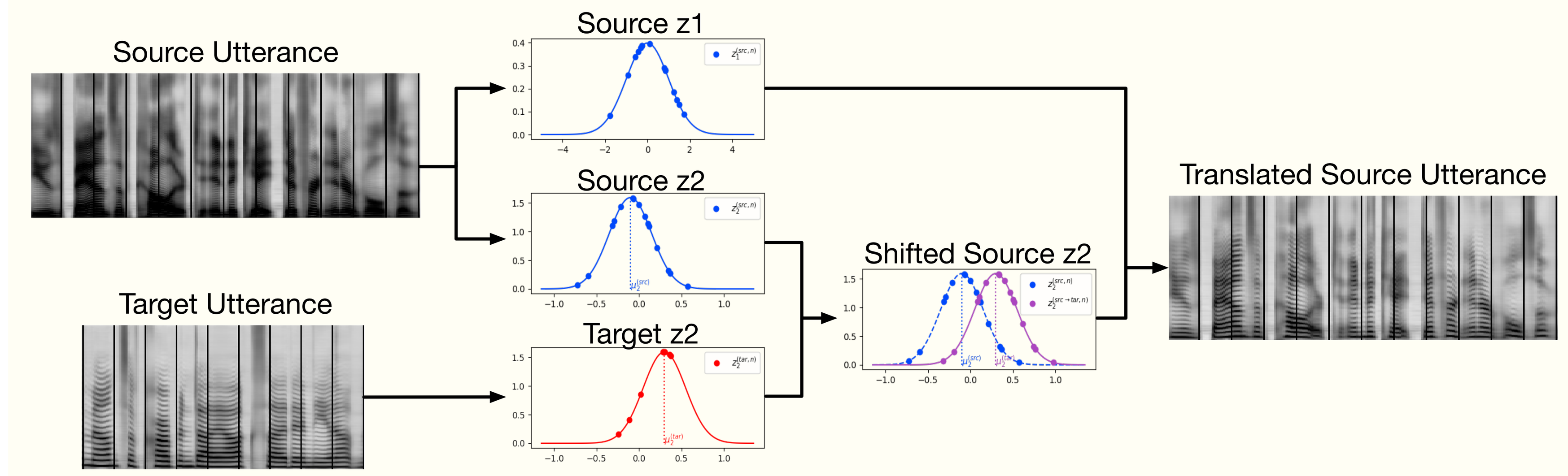
- **C** should preserve **A**'s segment-level attributes, such as phonetic content, and
- **C** should exhibit **B**'s sequence-level attributes, such as speaker identity and volume
  - consistent linguistic content (contour of formants) within a row
  - consistent speaker identity (spacing between harmonics) within a column



Qualitative Evaluation – Audio Translation

We cast *denoising* and *voice conversion* as **audio translation** problems, which aim to transform sequence-level attributes while preserving segment-level attributes.

- In our framework, it is equivalent to mapping the distribution of latent sequence variables of the source utterance $\boldsymbol{X}^{(src)}$ to that of the target utterance $\boldsymbol{X}^{(tar)}$.
- For each segment in $\boldsymbol{X}^{(src)}$, shift $\boldsymbol{z}_2^{(src,n)}$ by $\Delta\mu_2 = \mu_2^{(tar)} - \mu_2^{(src)}$. Keep $\boldsymbol{z}_1^{(src,n)}$ unaltered.
  - sequence-level attributes (volume/pitch) are translated, while linguistic content is preserved.
  - relative volume levels between segments in the source sequence are preserved.



Quantitative Evaluation – Speaker Verification

S-vectors and **latent sequence variables** should capture information about sequence-level attributes. We evaluate this property quantitatively via a **speaker verification** task on TIMIT: (full table is available in paper)

Features	Dimension	α	Raw	LDA (12 dim)	LDA (24 dim)
i-vector	48	-	10.12%	6.25%	5.95%
	100	-	9.52%	6.10%	5.50%
	200	-	9.82%	6.54%	6.10%
μ_2	16	0	5.06%	4.02%	-
	16	10	**2.38%**	**2.08%**	-
	32	10	**2.38%**	**2.08%**	**1.34%**
μ_1	16	10	27.68%	22.17%	-
	32	10	22.47%	16.82%	17.26%

Table: Comparison of speaker verification equal error rate (EER) on the TIMIT test set

Quantitative Evaluation – Domain Invariant ASR

We want to examine if **latent segment variables** contain segment-level attributes, *phonetic content*, but not sequence-level attributes, *speaker/environmental noise*.

- train an automatic speech recognition system using latent segment variables on one domain, and test on mismatched domains.

Train Set and Configuration			Test PER by Set		
ASR	FHVAE	Features	Male	Female	All
Train Male	-	FBank	**21.0%**	32.8%	25.2%
	Train All, $\alpha = 10$	z_1	22.0%	**26.2%**	**23.5%**

Table: TIMIT test phone error rate of acoustic models trained on different features and sets

Train Set and Configuration			Test WER by Set				
ASR	{FH-,β-}VAE	Features	Clean	Noisy	Channel	NC	All
Train Clean	-	FBank	**3.47%**	50.97%	36.99%	71.80%	55.51%
	Dev	z (VAE)	4.95%	23.54%	31.12%	46.21%	32.47%
	Dev, $\beta = 4$	z (β-VAE)	3.89%	24.40%	29.80%	47.87%	33.38%
	Dev, $\alpha = 10$	z_1 (FHVAE)	5.01%	**16.42%**	**20.29%**	**36.33%**	**24.41%**
	Dev, $\alpha = 10$	z_2 (FHVAE)	41.08%	68.73%	61.89%	86.36%	72.53%

Table: Aurora-4 test word error rate of acoustic models trained on different features and sets

wnhsu@mit.edu