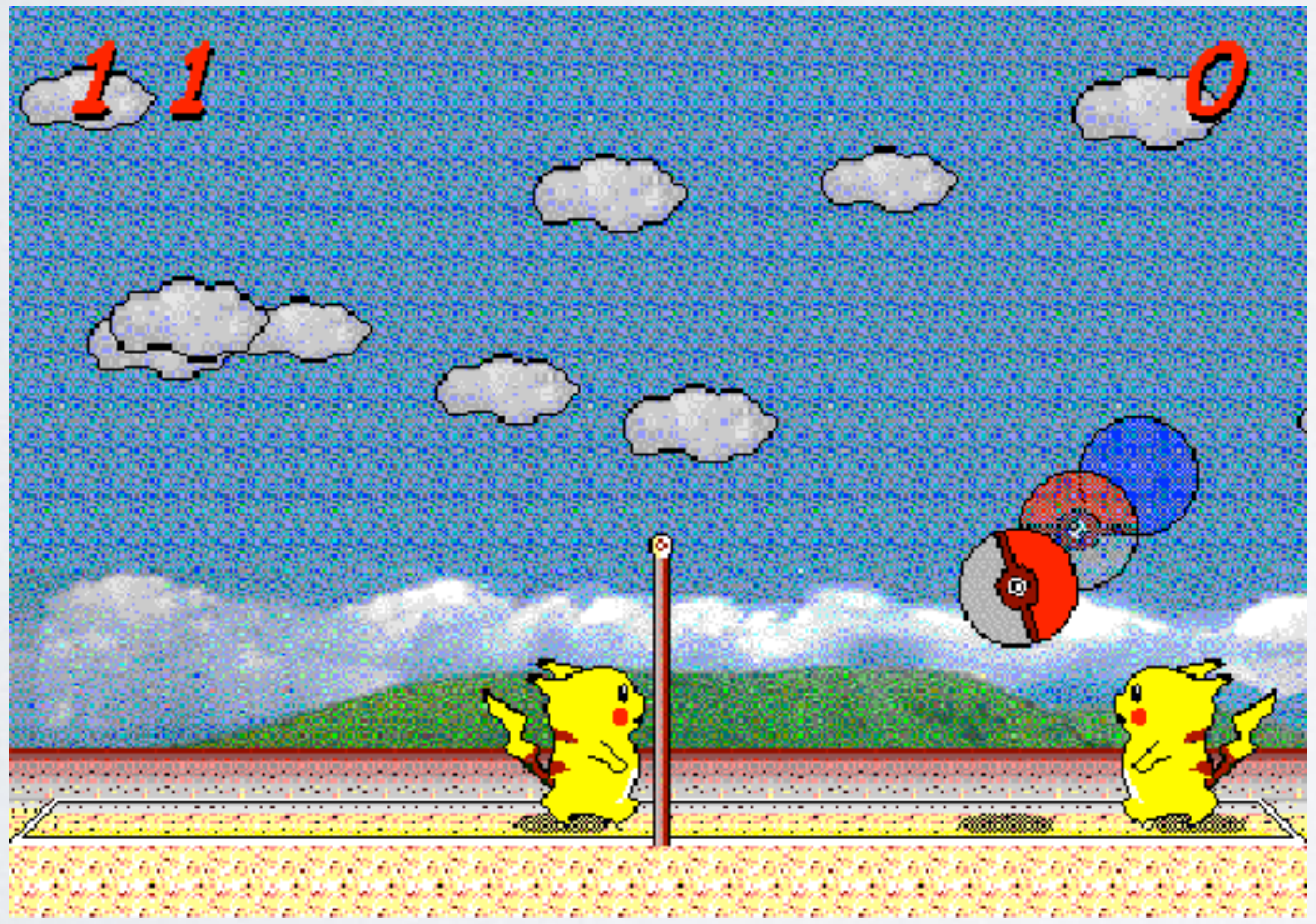
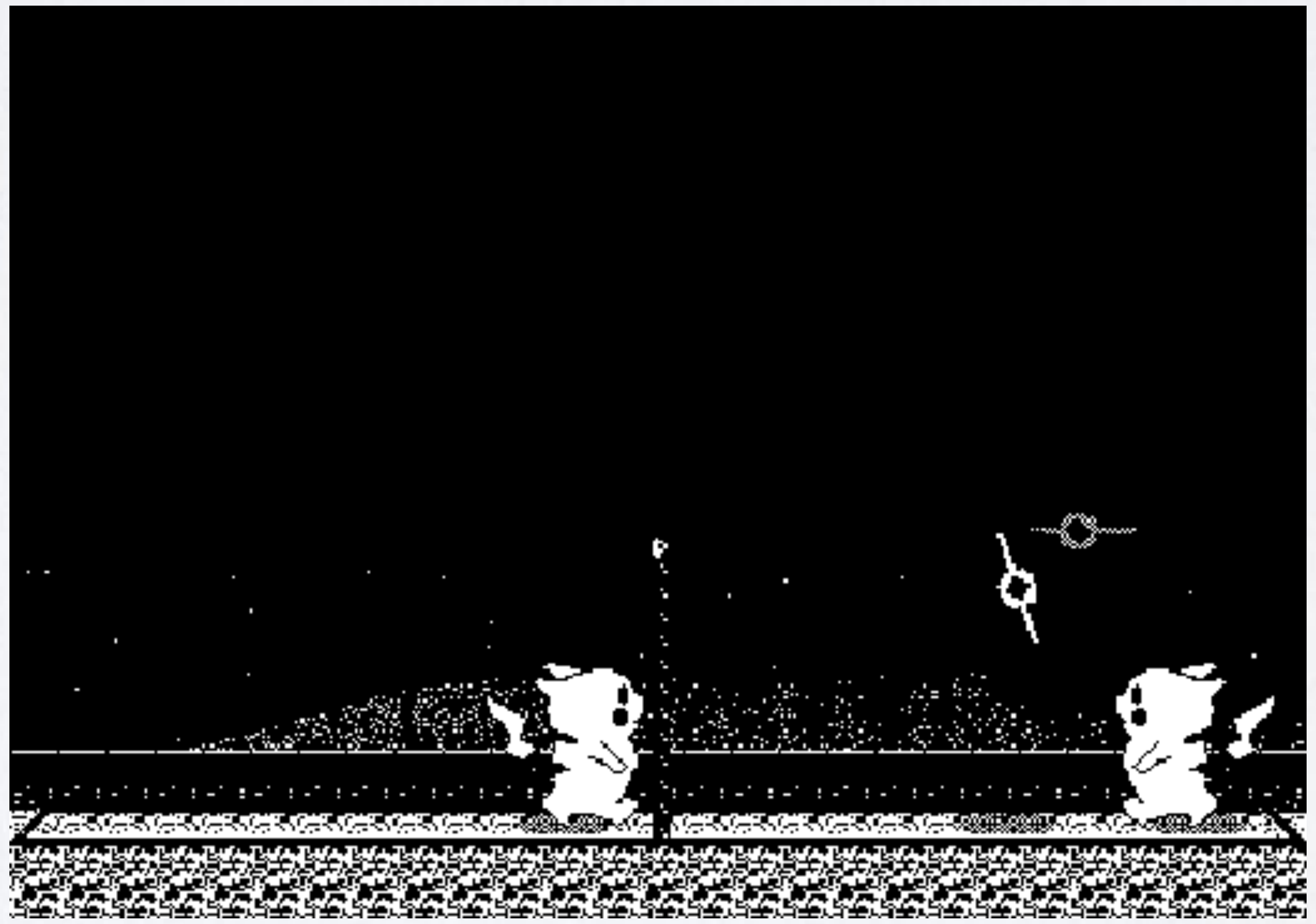
DISENTANGLING FROM SEQUENTIAL DATA

SNU Datamining Laboratory
2018. 6. 18 Seminar
Sungwon, Lyu
lyusungwon@dm.snu.ac.kr

REPRESENTATION LEARNING

- Representation Learning
  - Learning representation of the data that make it easier to extract useful information when building classifiers or other predictors





(110, 5, 225, 5, 210, 10)

Source: Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

WHAT MAKES A REPRESENTATION GOOD?

- Smoothness: $x \approx y$ then $f(x) \approx f(y)$
- Multiple Explanatory factors: Generalize many configurations of factors
- A hierarchical organization of explanatory factors
- Semi-supervised learning
- Shared factors across tasks
- Manifolds: Low meaningful dimensions
- Natural clustering: Named, categorized
- Temporal and Spatial coherence
- Sparsity: insensitive to small variations of x

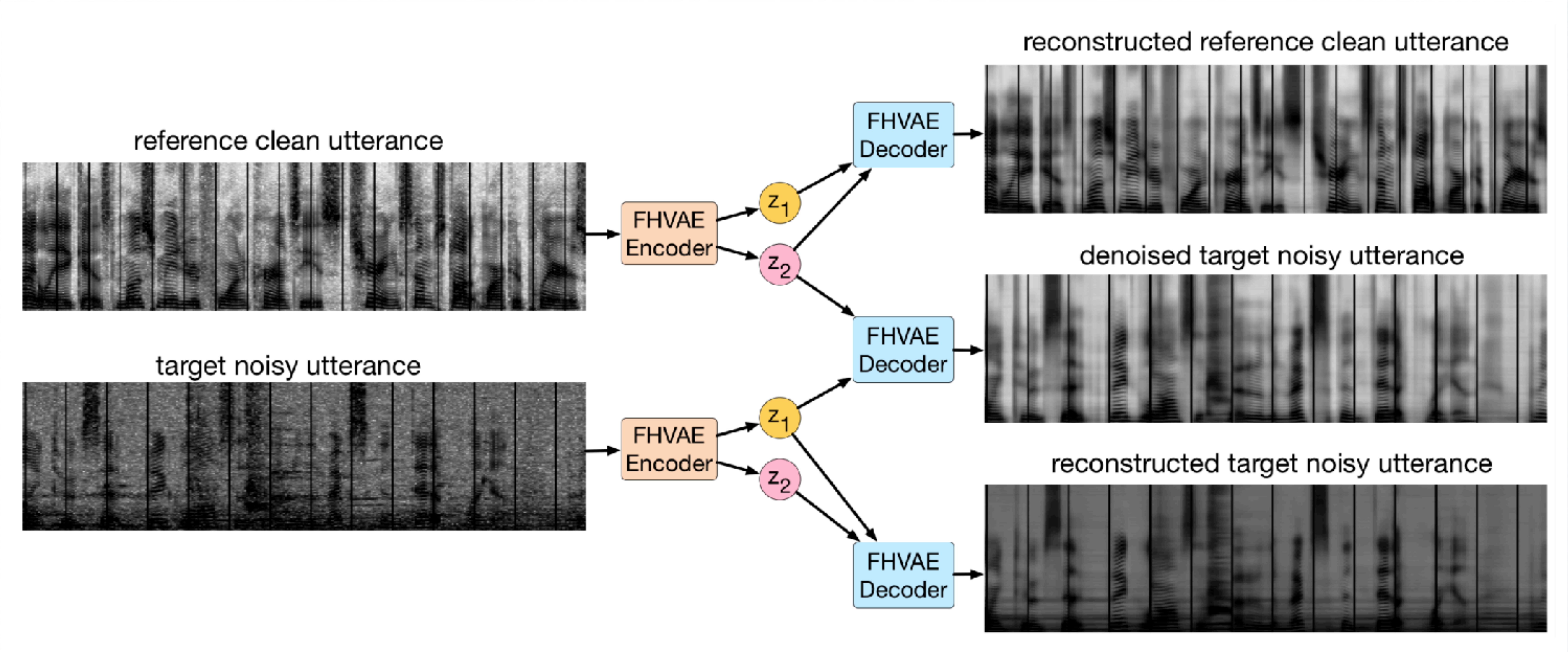
Source: Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

DISENTANGLING

- DC-IGN
  - Clamping a part of the hidden units for a pair of data points that are known to match in all but one factors of variation
- InfoGAN
  - Maximize the mutual information between a latent code c and x through the use of an auxiliary distribution Q(c|x)
- Beta-Vae
  - Encourages the latent representation to be factorised by adding beta to VAE objective

Source: Kulkarni, Tejas D., et al. "Deep convolutional inverse graphics network." *Advances in Neural Information Processing Systems*. 2015., Chen, Xi, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." *Advances in Neural Information Processing Systems*. 2016., Kulkarni, Tejas D., et al. "Deep convolutional inverse graphics network." *Advances in Neural Information Processing Systems*. 2015.

MOTIVATION

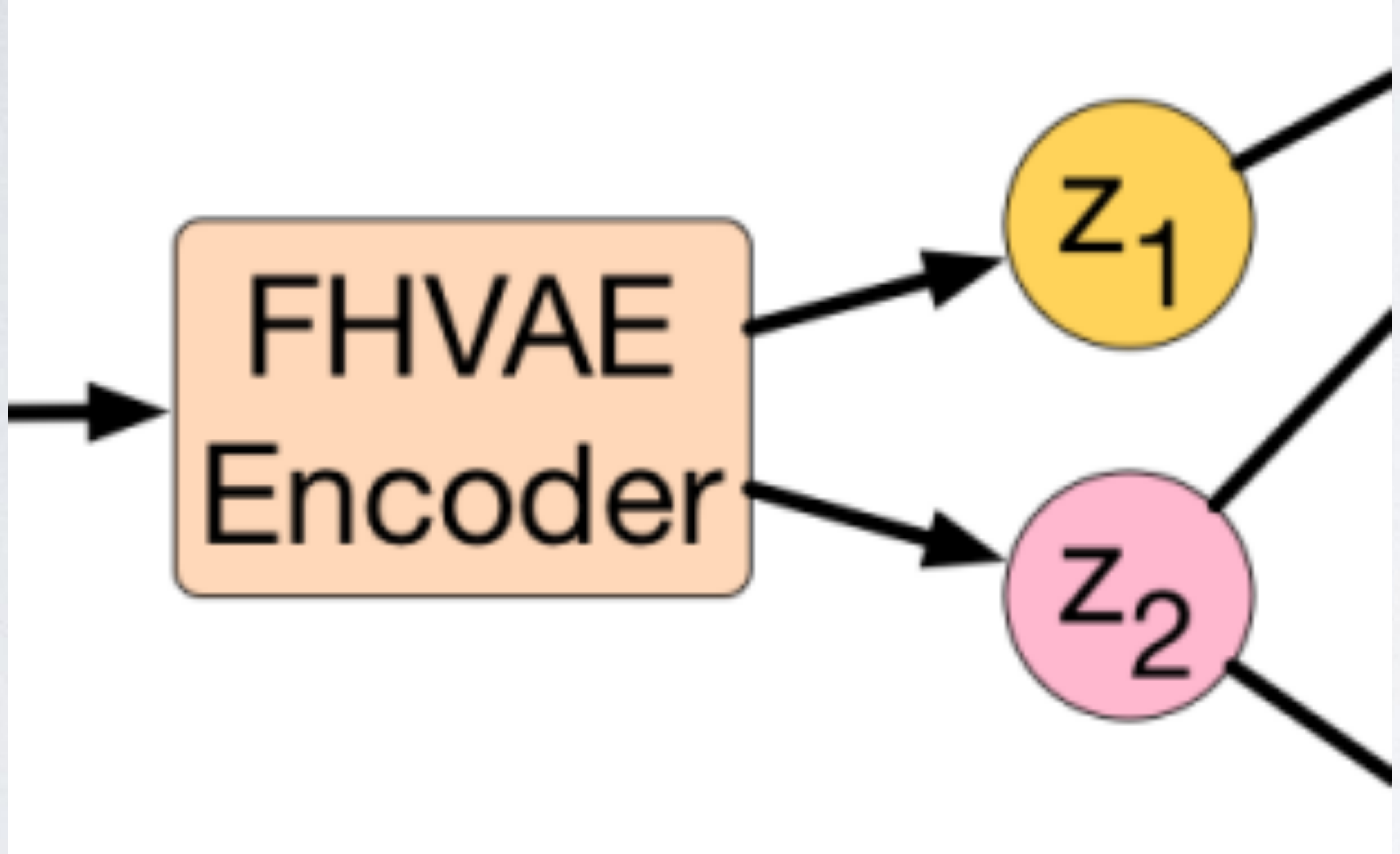


https://www.youtube.com/watch?v=naJZITvCfl4&feature=youtu.be

https://www.youtube.com/watch?v=VMX3IZYWYdg&feature=youtu.be

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

FHVAE



- Factorized Hierarchical VAE
- Sequence-level attributes
  - Speaker identity, Character style
  - Latent sequence variable (Z_1)
- Segment-level attributes
  - Phonetic content, Action
  - Latent segment variable (Z_2)
- Notation

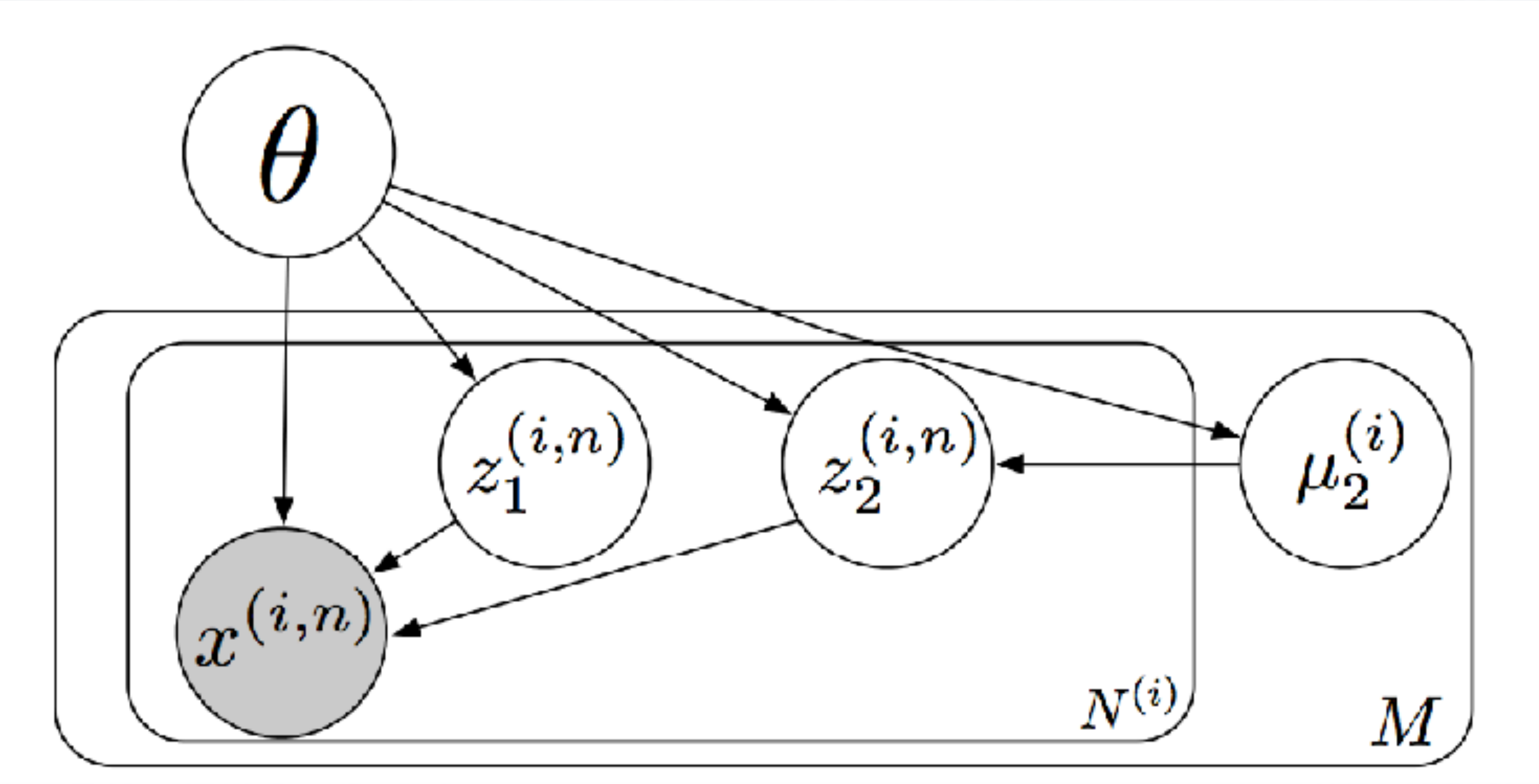
$$\mathcal{D} = \{\mathbf{X}^{(i)}\}_{i=1}^{M}, \mathbf{X}^{(i)} = \{\mathbf{x}^{(i,n)}\}_{n=1}^{N^{(i)}}$$

$$\mathbf{Z}_1 = \{\mathbf{z}_1^{(\mathbf{n})}\}, \mathbf{Z}_2 = \{\mathbf{z}_2^{(\mathbf{n})}\}$$

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

FHVAE

- Generative model



- Equation

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

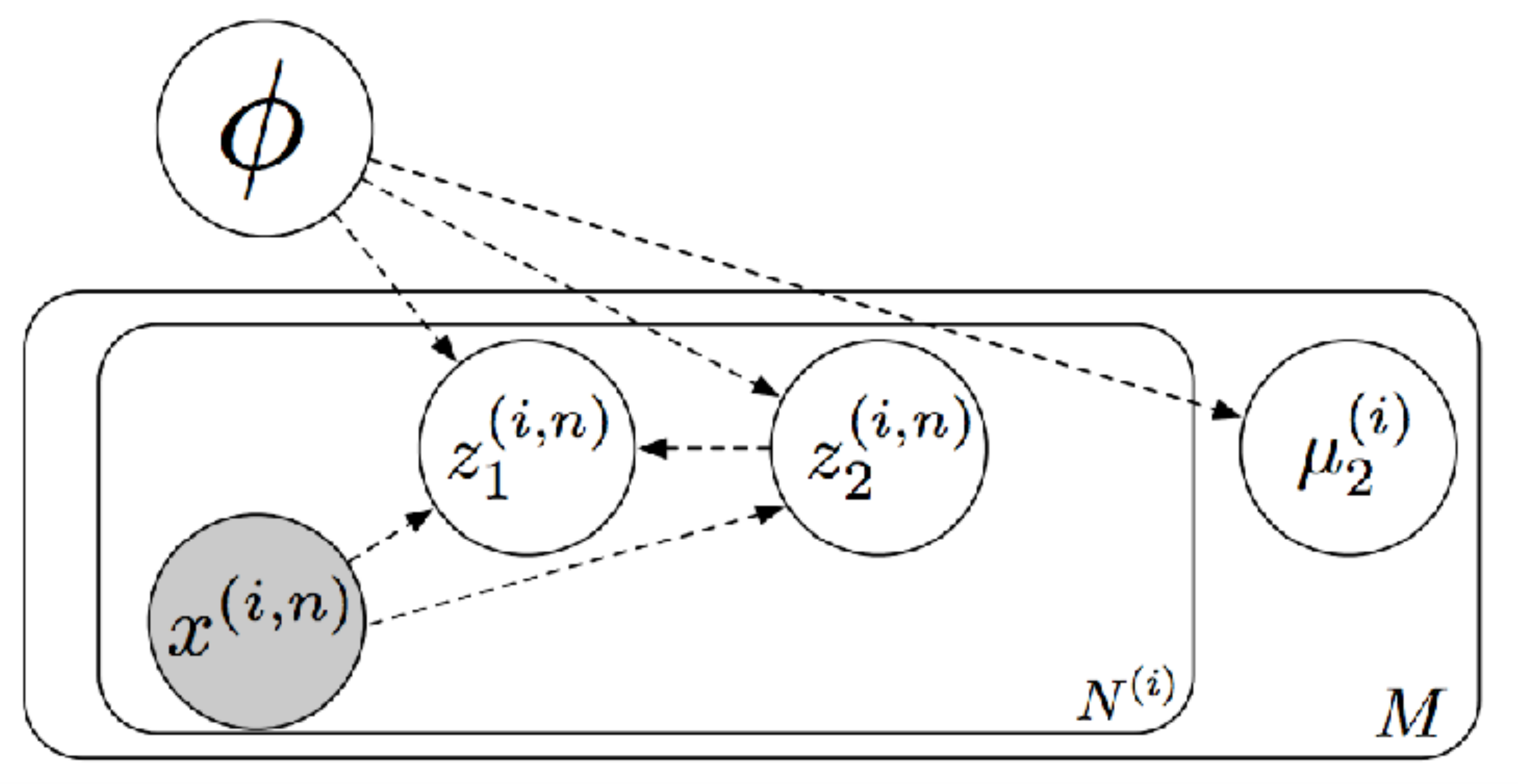
$$p_\theta(\boldsymbol{x} | \boldsymbol{z}_1, \boldsymbol{z}_2) = \mathcal{N}(\boldsymbol{x} | f_{\mu_x}(\boldsymbol{z}_1, \boldsymbol{z}_2), diag(f_{\sigma_x^2}(\boldsymbol{z}_1, \boldsymbol{z}_2)))$$

$$p_\theta(\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_1 | \boldsymbol{0}, \sigma_{\boldsymbol{z}_1}^2 \boldsymbol{I}), \quad p_\theta(\boldsymbol{z}_2 | \boldsymbol{\mu}_2) = \mathcal{N}(\boldsymbol{z}_2 | \boldsymbol{\mu}_2, \sigma_{\boldsymbol{z}_2}^2 \boldsymbol{I}), \quad p_\theta(\boldsymbol{\mu}_2) = \mathcal{N}(\boldsymbol{\mu}_2 | \boldsymbol{0}, \sigma_{\boldsymbol{\mu}_2}^2 \boldsymbol{I})$$

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

FHVAE

- Inference model



- Equation

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

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

$$q_\phi(\boldsymbol{z}_1 | \boldsymbol{x}, \boldsymbol{z}_2) = \mathcal{N}(\boldsymbol{z}_1 | g_{\mu_{\boldsymbol{z}_1}}(\boldsymbol{x}, \boldsymbol{z}_2), diag(g_{\sigma^2_{\boldsymbol{z}_1}}(\boldsymbol{x}, \boldsymbol{z}_2))),$$

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

FHVAE

- Factorising variational lower bound

$$\mathcal{L}(\theta,\phi;\boldsymbol{X}) = \sum_{n=1}^{N} \mathcal{L}(\theta,\phi;\boldsymbol{x}^{(n)}|\tilde{\boldsymbol{\mu}}_2) + \log p_\theta(\tilde{\boldsymbol{\mu}}_2) + const$$

$$\begin{aligned}\mathcal{L}(\theta,\phi;\boldsymbol{x}^{(n)}|\tilde{\boldsymbol{\mu}}_2) =& \mathbb{E}_{q_\phi(\boldsymbol{z}_1^{(n)},\boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})}\big[\log p_\theta(\boldsymbol{x}^{(n)}|\boldsymbol{z}_1^{(n)},\boldsymbol{z}_2^{(n)})\big] \\ &- \mathbb{E}_{q_\phi(\boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})}\big[D_{KL}(q_\phi(\boldsymbol{z}_1^{(n)}|\boldsymbol{x}^{(n)},\boldsymbol{z}_2^{(n)})||p_\theta(\boldsymbol{z}_1^{(n)}))\big] \\ &- D_{KL}(q_\phi(\boldsymbol{z}_2^{(n)}|\boldsymbol{x}^{(n)})||p_\theta(\boldsymbol{z}_2^{(n)}|\tilde{\boldsymbol{\mu}}_2)).\end{aligned}$$

- Segment Variational lower bound

$$\mathcal{L}(\theta,\phi;\boldsymbol{x}^{(n)}) = \mathcal{L}(\theta,\phi;\boldsymbol{x}^{(n)}|\tilde{\boldsymbol{\mu}}_2) + \frac{1}{N}\log p_\theta(\tilde{\boldsymbol{\mu}}_2) + const.$$

- Discriminative objective

$$\begin{aligned}\log p(i|\boldsymbol{z}_2^{(i,n)}) &= \log p(\boldsymbol{z}_2^{(i,n)}|i) - \log\sum_{j=1}^{M} p(\boldsymbol{z}_2^{(i,n)}|j) \quad (p(i) \text{ is assumed uniform}) \\ &:= \log p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\boldsymbol{\mu}}_2^{(i)}) - \log\big(\sum_{j=1}^{M} p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\boldsymbol{\mu}}_2^{(j)})\big),\end{aligned}$$

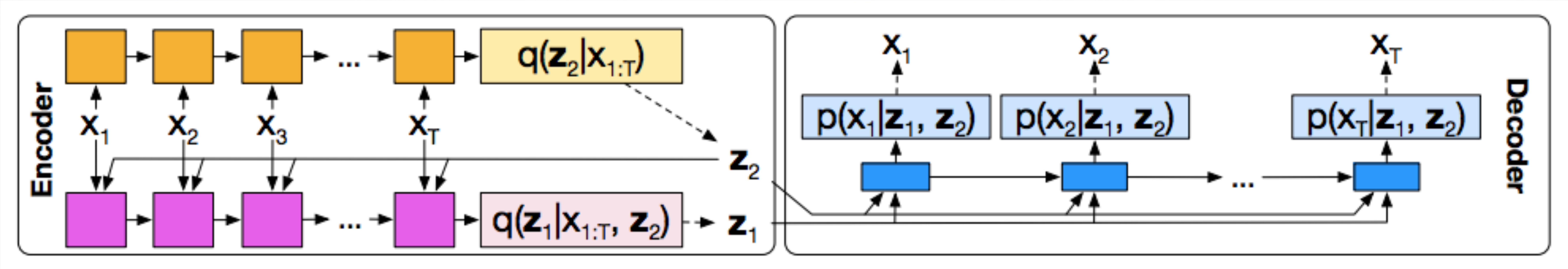
- Objetive function

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

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

IMPLEMENTATION

- Sequence to sequence autoencoder model



$$
\begin{aligned}
(\boldsymbol{h}_{\boldsymbol{z}_2,t}, \boldsymbol{c}_{\boldsymbol{z}_2,t}) &= \mathrm{LSTM}(x_{t-1}, \boldsymbol{h}_{\boldsymbol{z}_2,t-1}, \boldsymbol{c}_{\boldsymbol{z}_2,t-1}; \phi_{\mathrm{LSTM},\boldsymbol{z}_2}) \\
q_\phi(\boldsymbol{z}_2|x_{1:T}) &= \mathcal{N}(\boldsymbol{z}_2|\,\mathrm{MLP}(\boldsymbol{h}_{\boldsymbol{z}_2,T}; \phi_{\mathrm{MLP}_{\mu},\boldsymbol{z}_2}), \mathrm{diag}(\exp(\mathrm{MLP}(\boldsymbol{h}_{\boldsymbol{z}_2,T}; \phi_{\mathrm{MLP}_{\sigma^2},\boldsymbol{z}_2})))) \\
(\boldsymbol{h}_{\boldsymbol{z}_1,t}, \boldsymbol{c}_{\boldsymbol{z}_1,t}) &= \mathrm{LSTM}([x_{t-1}; \boldsymbol{z}_2], \boldsymbol{h}_{\boldsymbol{z}_1,t-1}, \boldsymbol{c}_{\boldsymbol{z}_1,t-1}; \phi_{\boldsymbol{z}_1}) \\
q_\phi(\boldsymbol{z}_1|x_{1:T}, \boldsymbol{z}_2) &= \mathcal{N}(\boldsymbol{z}_1|\,\mathrm{MLP}(\boldsymbol{h}_{\boldsymbol{z}_1,T}; \phi_{\mathrm{MLP}_{\mu},\boldsymbol{z}_1}), \mathrm{diag}(\exp(\mathrm{MLP}(\boldsymbol{h}_{\boldsymbol{z}_1,T}; \phi_{\mathrm{MLP}_{\sigma^2},\boldsymbol{z}_1})))) \\
(\boldsymbol{h}_{\boldsymbol{x},t}, \boldsymbol{c}_{\boldsymbol{x},t}) &= \mathrm{LSTM}([\boldsymbol{z}_1; \boldsymbol{z}_2], \boldsymbol{h}_{\boldsymbol{x},t-1}, \boldsymbol{c}_{\boldsymbol{x},t-1}; \phi_{\boldsymbol{x}}) \\
p_\theta(x_t|\boldsymbol{z}_1, \boldsymbol{z}_2) &= \mathcal{N}(\boldsymbol{x}_t|\,\mathrm{MLP}(\boldsymbol{h}_{\boldsymbol{x},t}; \phi_{\mathrm{MLP}_{\mu},\boldsymbol{x}}), \mathrm{diag}(\exp(\mathrm{MLP}(\boldsymbol{h}_{\boldsymbol{x},t}; \phi_{\mathrm{MLP}_{\sigma^2},\boldsymbol{x}})))),
\end{aligned}
$$

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

EXPERIMENT

- Speaker Verification

Table 1: Comparison of speaker verification equal error rate (EER) on the TIMIT test set

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%
$\boldsymbol{\mu}_2$	16	0	5.06%	4.02%	-
	16	10^{-1}	4.91%	4.61%	-
	16	10^0	3.87%	3.86%	-
	16	10^1	**2.38%**	**2.08%**	-
	32	10^1	**2.38%**	**2.08%**	**1.34%**
$\boldsymbol{\mu}_1$	16	10^0	22.77%	15.62%	-
	16	10^1	27.68%	22.17%	-
	32	10^1	22.47%	16.82%	17.26%

- Automatic Speech Recognition

Table 2: TIMIT test phone error rate of acoustic models trained on different features and sets

Train Set and Configuration			Test PER by Set		
ASR	FHVAE	Features	Male	Female	All
Train All	-	FBank	20.1%	16.7%	19.1%
Train Male	-	FBank	**21.0%**	32.8%	25.2%
	Train All, $\alpha = 10$	$\boldsymbol{z}_1$	22.0%	**26.2%**	**23.5%**

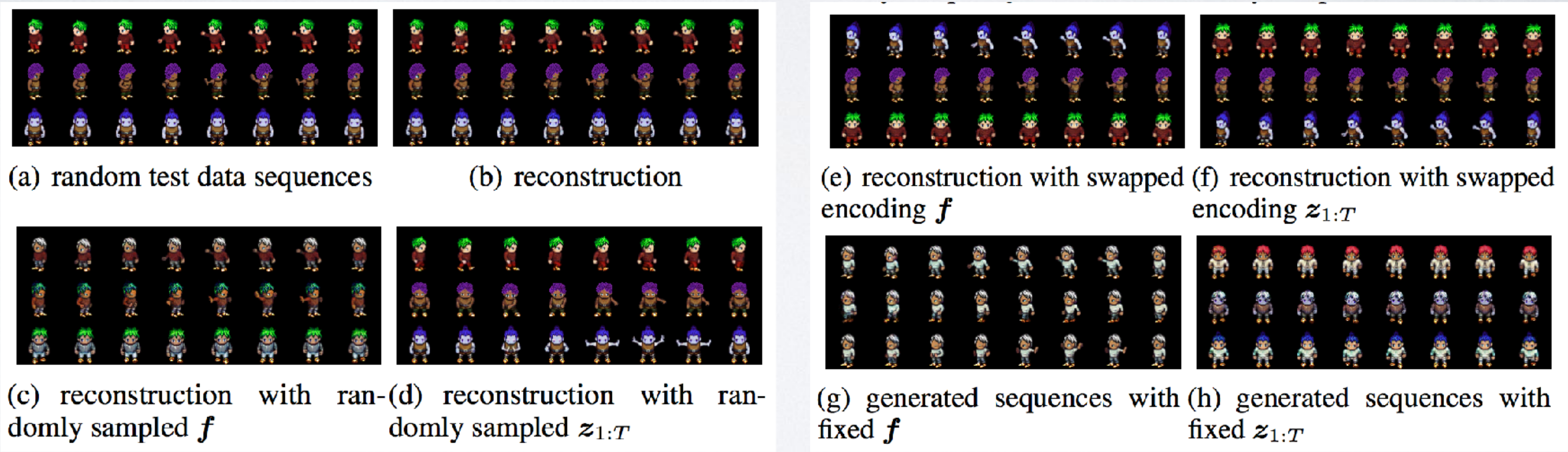
Table 3: Aurora-4 test word error rate of acoustic models trained on different features and sets

Train Set and Configuration			Test WER by Set				
ASR	{FH-,β-}VAE	Features	Clean	Noisy	Channel	NC	All
Train All	-	FBank	3.60%	7.06%	8.24%	18.49%	11.80%
Train Clean	-	FBank	**3.47%**	50.97%	36.99%	71.80%	55.51%
	Dev, $\beta = 1$	$\boldsymbol{z}$ (β-VAE)	4.95%	23.54%	31.12%	46.21%	32.47%
	Dev, $\beta = 2$	$\boldsymbol{z}$ (β-VAE)	3.57%	27.24%	30.56%	48.17%	34.75%
	Dev, $\beta = 4$	$\boldsymbol{z}$ (β-VAE)	3.89%	24.40%	29.80%	47.87%	33.38%
	Dev, $\beta = 8$	$\boldsymbol{z}$ (β-VAE)	5.32%	34.84%	36.13%	58.02%	42.76%
	Dev, $\alpha = 10$	$\boldsymbol{z}_1$ (FHVAE)	5.01%	**16.42%**	**20.29%**	**36.33%**	**24.41%**
	Dev, $\alpha = 10$	$\boldsymbol{z}_2$ (FHVAE)	41.08%	68.73%	61.89%	86.36%	72.53%
	Dev\NC, $\alpha = 10$	$\boldsymbol{z}_1$ (FHVAE)	5.25%	16.52%	19.30%	40.59%	26.23%

Hsu, Wei-Ning, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." *Advances in neural information processing systems*. 2017.

RELATED WORKS

- A Deep Generative Model for Disentangled Representations of Sequential Data
  - Applied almost the same architecture to video



(a) random test data sequences
(b) reconstruction
(c) reconstruction with randomly sampled $\boldsymbol{f}$
(d) reconstruction with randomly sampled $\boldsymbol{z}_{1:T}$
(e) reconstruction with swapped encoding $\boldsymbol{f}$
(f) reconstruction with swapped encoding $\boldsymbol{z}_{1:T}$
(g) generated sequences with fixed $\boldsymbol{f}$
(h) generated sequences with fixed $\boldsymbol{z}_{1:T}$

Source: Li, Yingzhen, and Stephan Mandt. "A Deep Generative Model for Disentangled Representations of Sequential Data." *arXiv preprint arXiv:1803.02991* (2018).

RELATED WORKS

- Multimodal Unsupervised Image-to-Image Translation
  - Used Adversarial training to disentangle style and content feature

(b) edges → shoes (b) big cats → house cats

Huang, Xun, et al. "Multimodal Unsupervised Image-to-Image Translation." *arXiv preprint arXiv:1804.04732* (2018).