PADERBORN UNIVERSITY
The University for the Information Society

Communications Engineering

NT

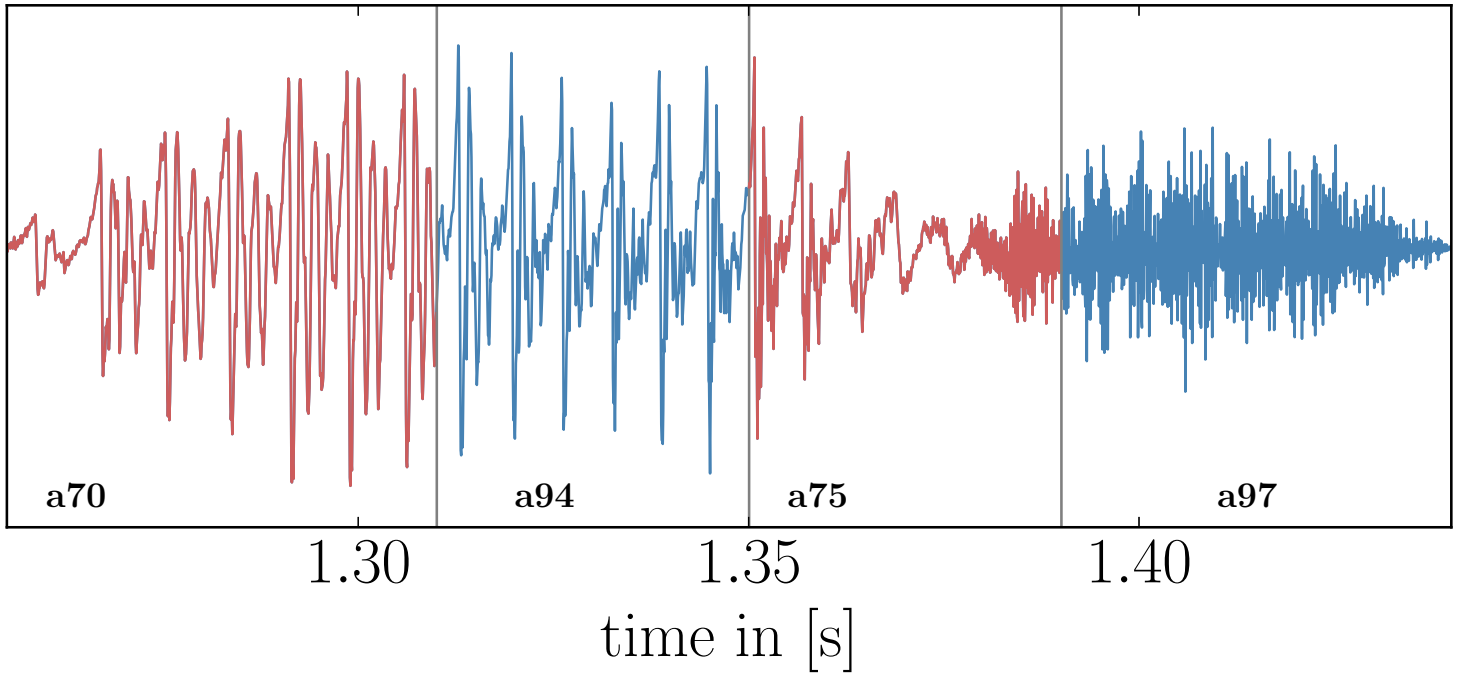
Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery

Janek Ebbers[1], Jahn Heymann[1], Lukas Drude[1], Thomas Glarner[1], Reinhold Haeb-Umbach[1], Bhiksha Raj[2]

[1]Department of Communications Engineering - Paderborn University
[2]Language Technologies Institute - Carnegie Mellon University

20.08.2017

Introduction (1)



Acoustic unit discovery (AUD)

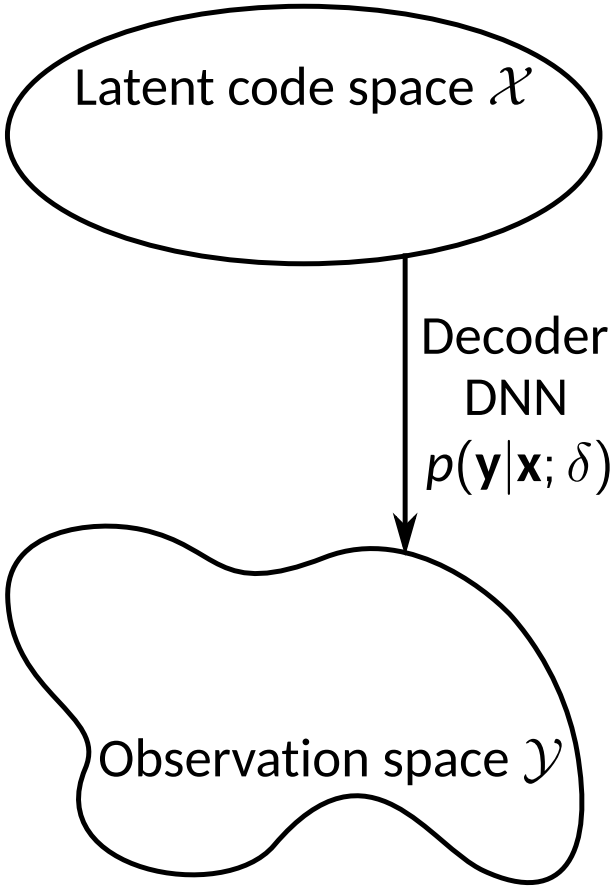
- Learning acoustic units (phonetic inventory) from raw speech
- Unsupervised training of generative model
- SOTA: GMM/HMM

Introduction (2)

Motivation

- Known from ASR: Superiority of DNNs over GMMs
  - But: Discriminative DNNs not transferable to AUD
- Variational Autoencoder (VAE)
  - Deep generative model
  - Sophisticated data distribution modeling by DNN
  - Efficient variational inference by DNN
- **Here:** Marrying VAE with HMM for AUD with sophisticated emission distribution modeling

Variational Autoencoder (1)



Model

- Latent codes:

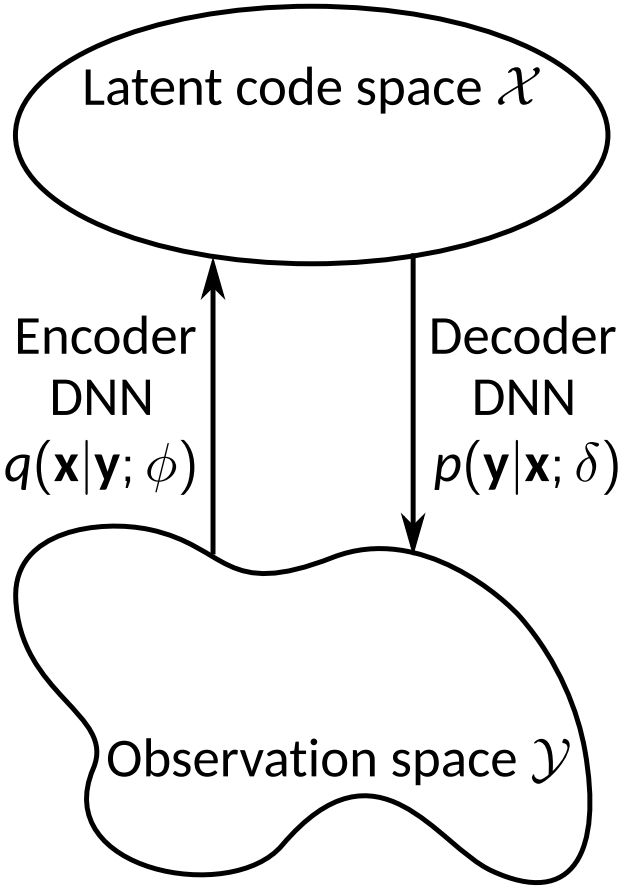
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I})$$

- Non-linear observation model:

$$\mathbf{y} = f(\mathbf{x}; \delta) + \mathbf{v}; \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

$$\Rightarrow p(\mathbf{y}|\mathbf{x}; \delta) = \mathcal{N}(\mathbf{y}; f(\mathbf{x}; \delta), \sigma^2\mathbf{I})$$

Variational Autoencoder (1)



Model

- Latent codes:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I})$$

- Non-linear observation model:

$$\mathbf{y} = f(\mathbf{x};\delta) + \mathbf{v}; \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

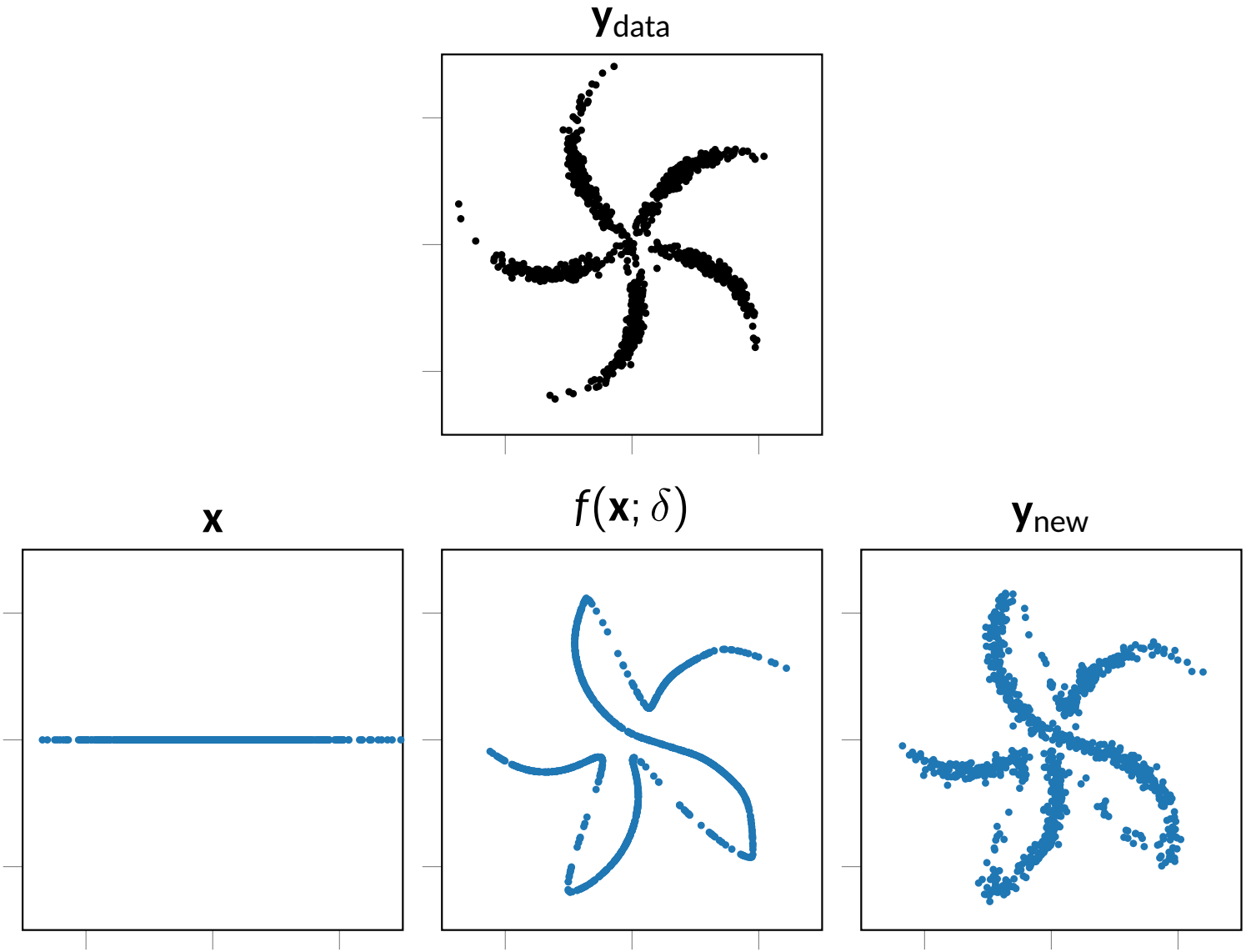
$$\Rightarrow p(\mathbf{y}|\mathbf{x};\delta) = \mathcal{N}(\mathbf{y}; f(\mathbf{x};\delta), \sigma^2\mathbf{I})$$

- Variational inference:

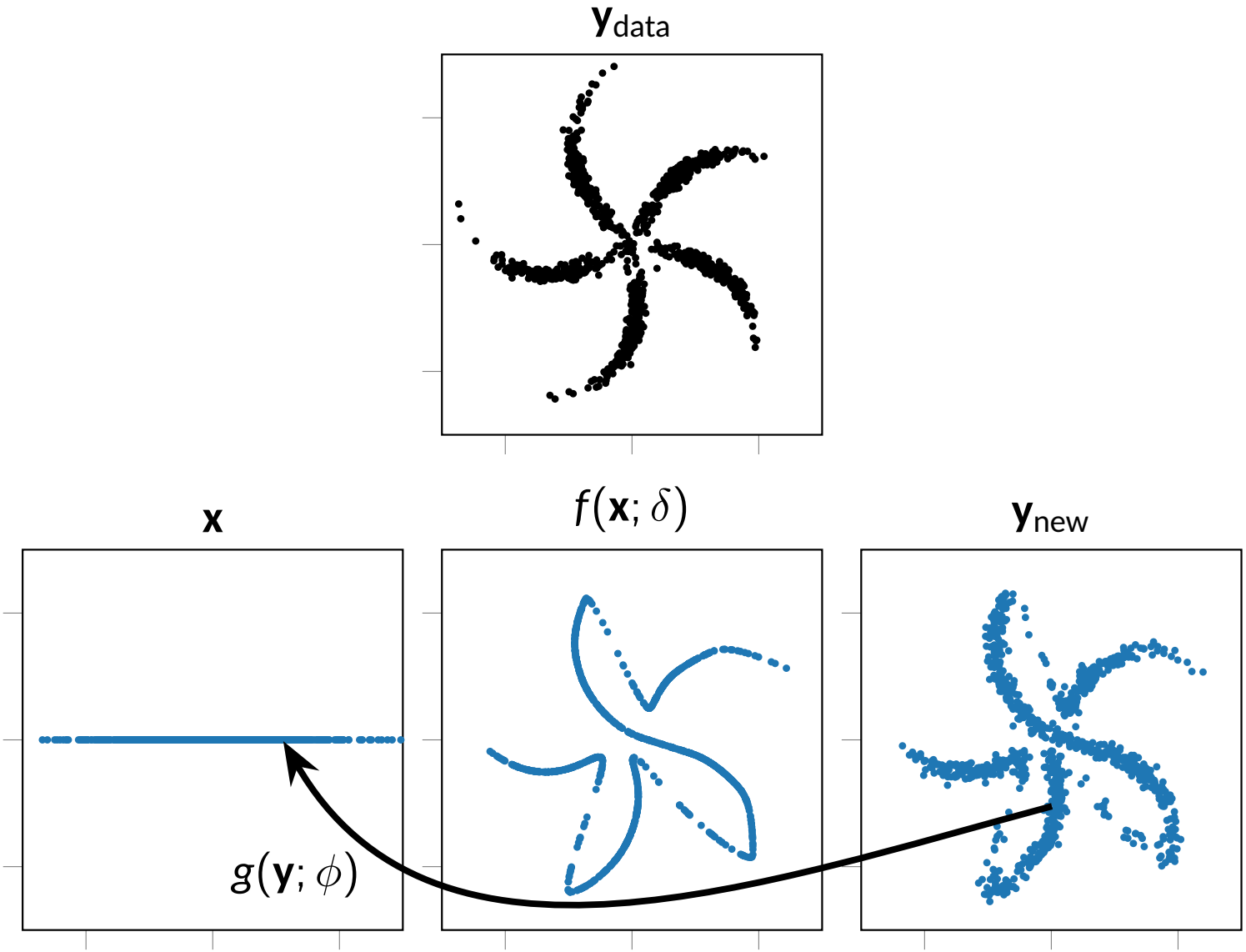
$$q(\mathbf{x}|\mathbf{y};\phi) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}})$$

$$(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \ln \boldsymbol{\sigma}_{\mathbf{x}|\mathbf{y}}) = g(\mathbf{y};\phi)$$

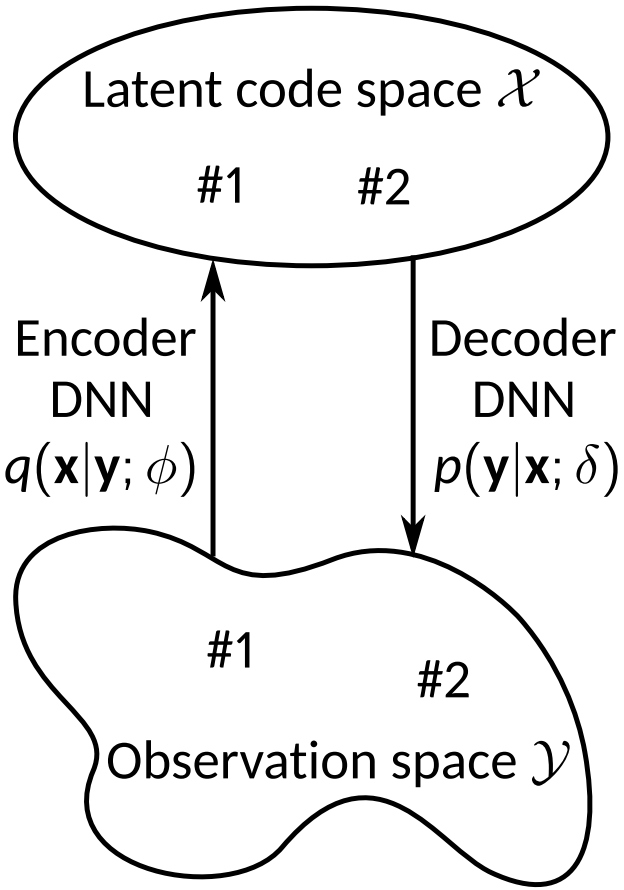
Variational Autoencoder (2)



Variational Autoencoder (2)



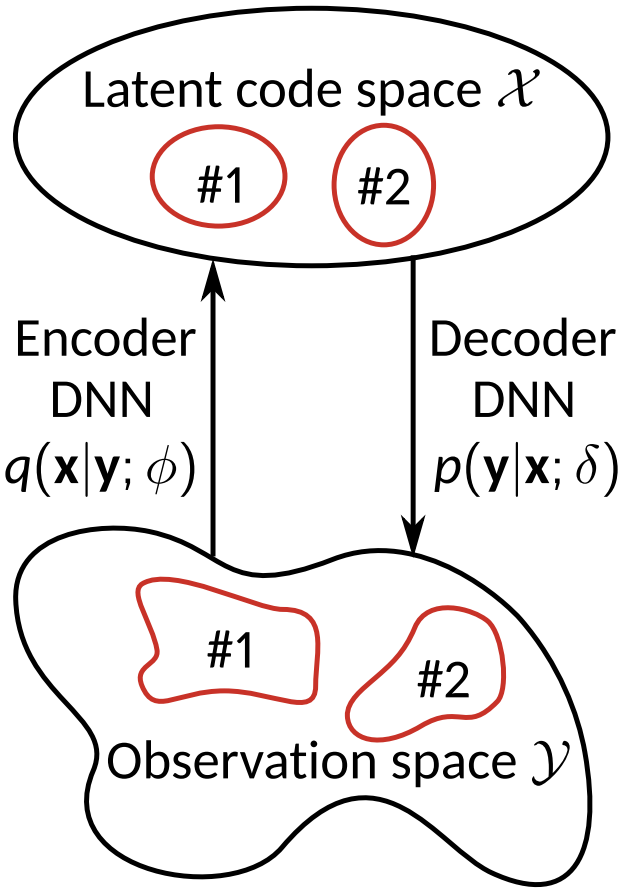
GMMVAE (1)



Model

- Latent codes:
  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I})$
- Non-linear observation model:
  $$\mathbf{y} = f(\mathbf{x};\delta) + \mathbf{v}; \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$
  $$\Rightarrow p(\mathbf{y}|\mathbf{x};\delta) = \mathcal{N}(\mathbf{y}; f(\mathbf{x};\delta), \sigma^2\mathbf{I})$$
- Variational inference:
  $$q(\mathbf{x}|\mathbf{y};\phi) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}})$$
  $$(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \ln \boldsymbol{\sigma}_{\mathbf{x}|\mathbf{y}}) = g(\mathbf{y};\phi)$$

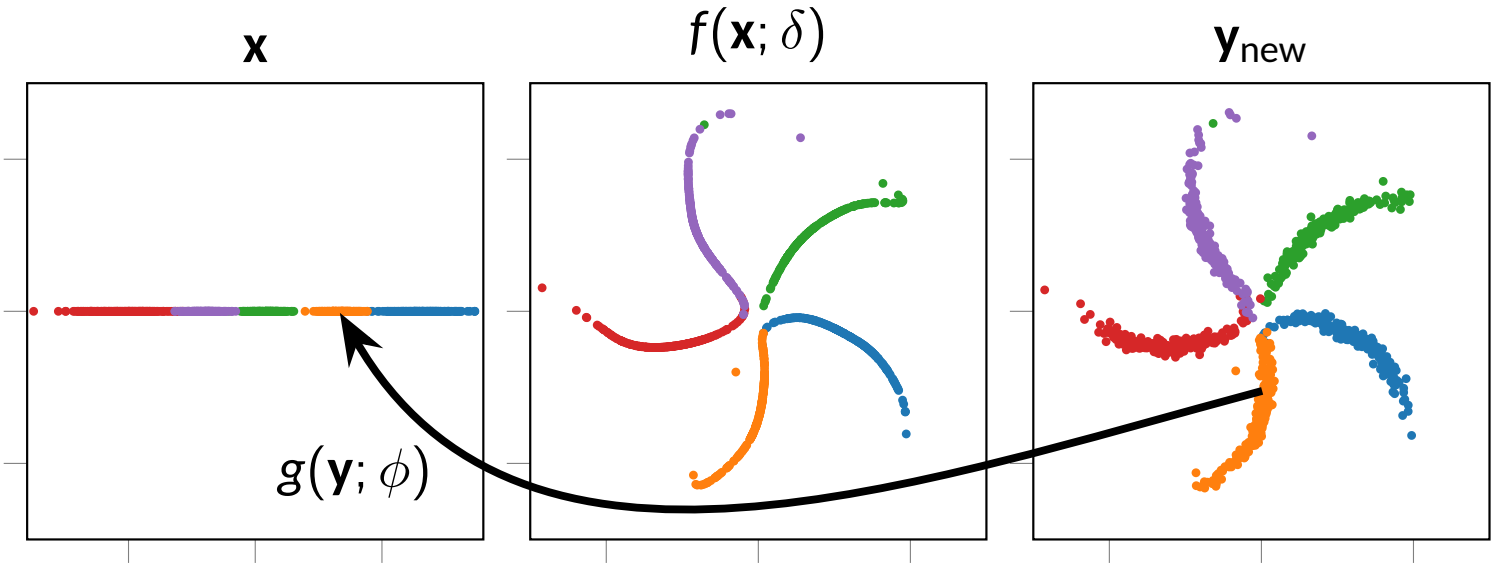
GMMVAE (1)



Model

- Latent codes:
  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}) \Rightarrow p(\mathbf{x}|z) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- Non-linear observation model:
  $$\mathbf{y} = f(\mathbf{x}; \delta) + \mathbf{v}; \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$
  $$\Rightarrow p(\mathbf{y}|\mathbf{x}; \delta) = \mathcal{N}(\mathbf{y}; f(\mathbf{x}; \delta), \sigma^2\mathbf{I})$$
- Variational inference:
  $$q(\mathbf{x}|\mathbf{y}; \phi) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}})$$
  $$(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \ln \boldsymbol{\sigma}_{\mathbf{x}|\mathbf{y}}) = g(\mathbf{y}; \phi)$$

GMMVAE (2)

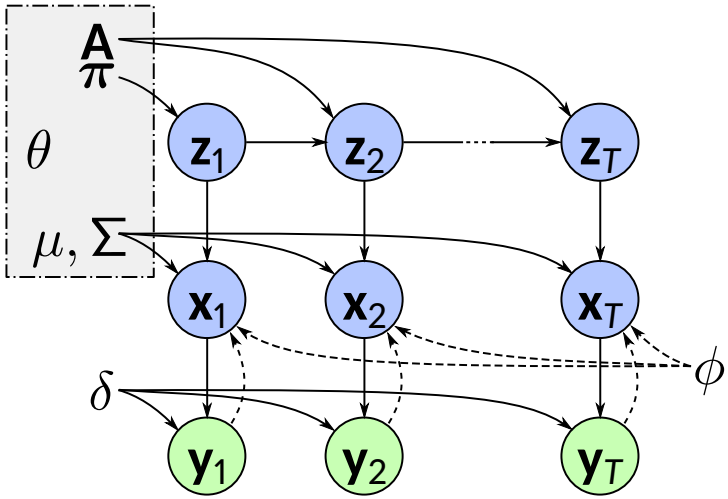


ML estimate

$$\hat{z} = \underset{z}{\operatorname{argmax}}\, b_z(\mathbf{y})$$

$$\ln b_z(\mathbf{y}) = -H\big(q(\mathbf{x}|\mathbf{y}), p(\mathbf{x}|z)\big) \quad \text{(acoustic score)}$$

HMMVAE



HMMVAE

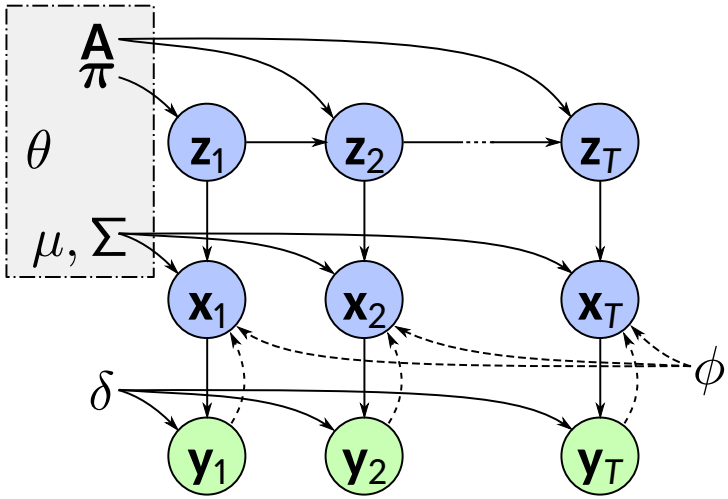
Inference

$$\ln b_z(\mathbf{y}) = -H\big(q(\mathbf{x}|\mathbf{y}), p(\mathbf{x}|z)\big)$$

$$q(Z|\mathbf{Y}) = \mathrm{FB}(\mathbf{Y}; b_z, \boldsymbol{\pi}, \mathbf{A})$$

$$\hat{Z} = \mathrm{Viterbi}(\mathbf{Y}; b_z, \boldsymbol{\pi}, \mathbf{A})$$

HMMVAE



HMMVAE

Inference

$$\ln b_z(\mathbf{y}) = -H\big(q(\mathbf{x}|\mathbf{y}), p(\mathbf{x}|z)\big)$$

$$q(Z|\mathbf{Y}) = \mathrm{FB}(\mathbf{Y}; b_z, \pi, \mathbf{A})$$

$$\hat{Z} = \mathrm{Viterbi}(\mathbf{Y}; b_z, \pi, \mathbf{A})$$

Objective

$$\mathcal{L}(\mathbf{Y}; \theta, \delta, \phi) = \underbrace{\mathbb{E}_{q(\mathbf{X}|\mathbf{Y};\phi)}\big[\ln p(\mathbf{Y}|\mathbf{X};\delta)\big]}_{\text{Reconstruction score}} - \underbrace{\mathrm{KL}\big(q(\mathbf{X},\mathbf{Z}|\mathbf{Y};\phi)||p(\mathbf{X},\mathbf{Z};\theta)\big)}_{\text{Regularization}}$$

Experiments

Task: Acoustic Unit Discovery (AUD)

- Database: Timit
- Unsupervised training of HMMVAE
- Segmentation of utterances

Model

- U=72 units, each modeled by three states (left-right)
- Features: 13 element MFCCs with Δ and $\Delta\Delta$
- Initialized using segmentation found by unsupervised GMM/HMM[1]

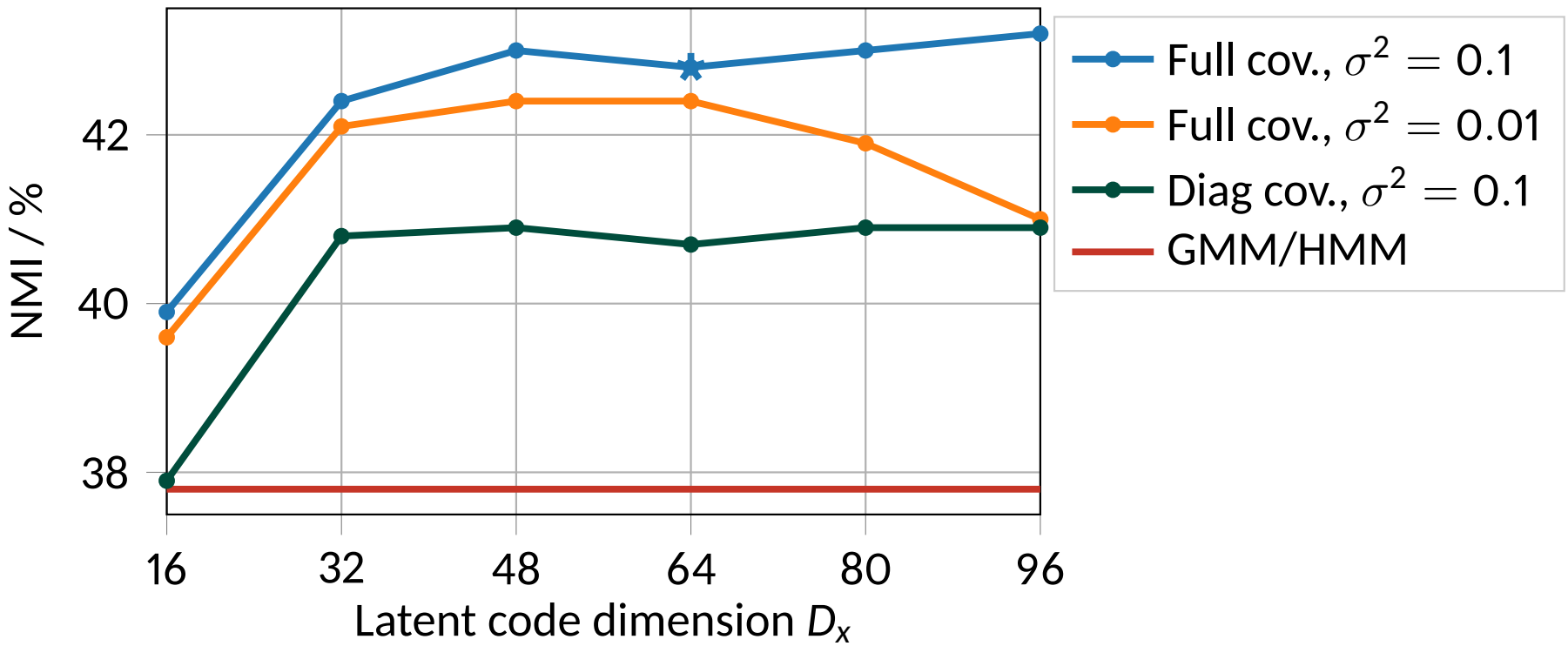
Performance measure

- Normalized mutual information (NMI)
- Equivalent phone error rate (eq. PER)

[1] L. Ondel, L. Burget, and J. Cernocky, "Variational Inference for Acoustic Unit Discovery"

Results

Model	Training	NMI	eq. PER
GMM/HMM	FB	37.8%	65.4%
HMMVAE	Viterbi	42.8%	58.9%
HMMVAE	FB	42.6%	59.0%



Conclusions

Summary

- Extended VAE by an HMM in latent code space to capture temporal correlations
- Derived iterative EM-like algorithm for inference and optimization
- Applied HMMVAE to unsupervised AUD task
- Significantly improved AUD performance over variational GMM/HMM in terms of NMI and eq. PER

Future Work

- Bayesian parameter estimation