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] Paderborn University, Germany [2]Carnegie Mellon University, United States

Introduction

- Acoustic unit discovery (AUD)
  - Learning AUs (phonetic inventory) from raw speech
  - Unsupervised training of generative model
  - SOTA: GMM/HMM
- Known from ASR: Superiority of DNNs over GMMs
  - But: Discriminative DNNs not transferable to AUD
- Variational Autoencoder (VAE) [1]
  - 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

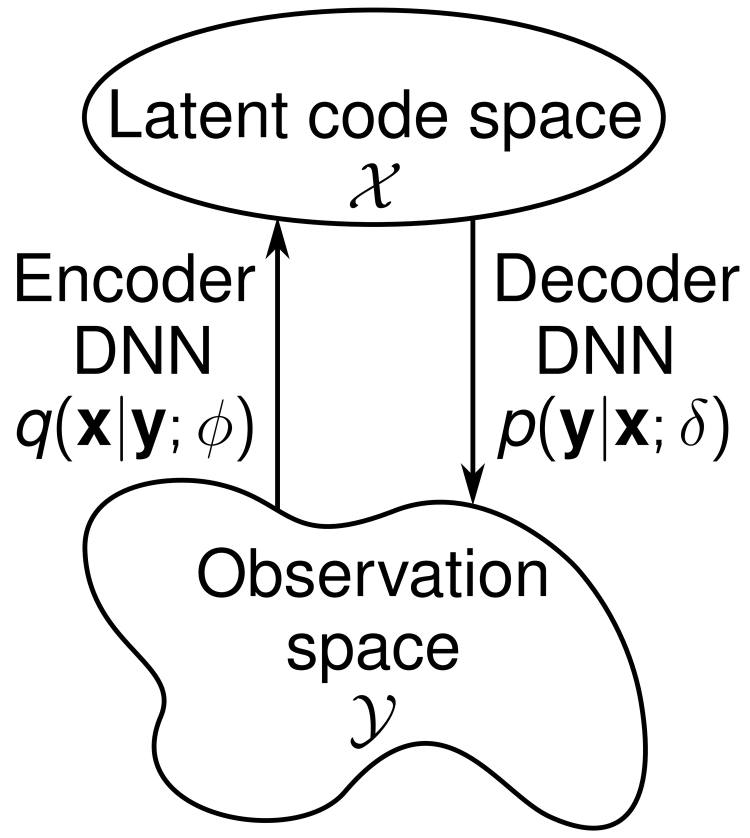
VAE

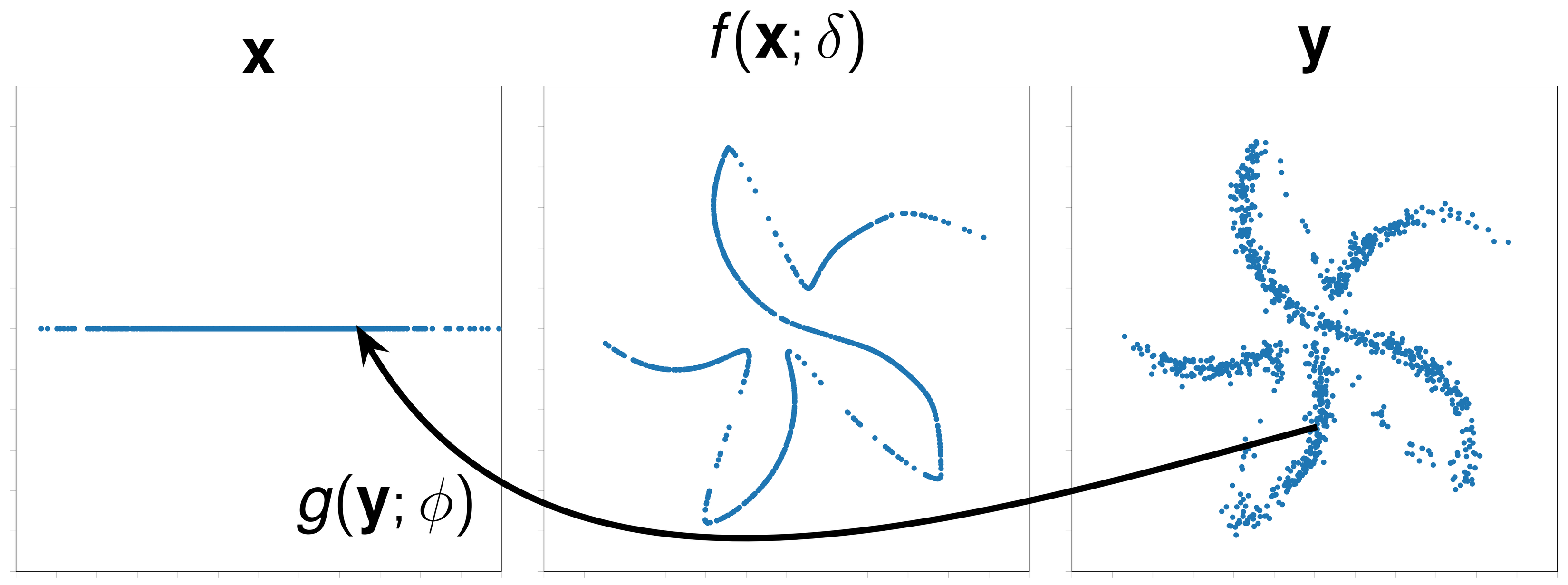
- Given: arbitrary distributed observations $\mathbf{y}$
- Assuming: simply structured latent codes $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- Modeling observations by decoder $f(\mathbf{x};\delta)$ and Gaussian observation noise $\mathbf{v}$:

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

- Variational inference by encoder $g(\mathbf{y};\phi)$:

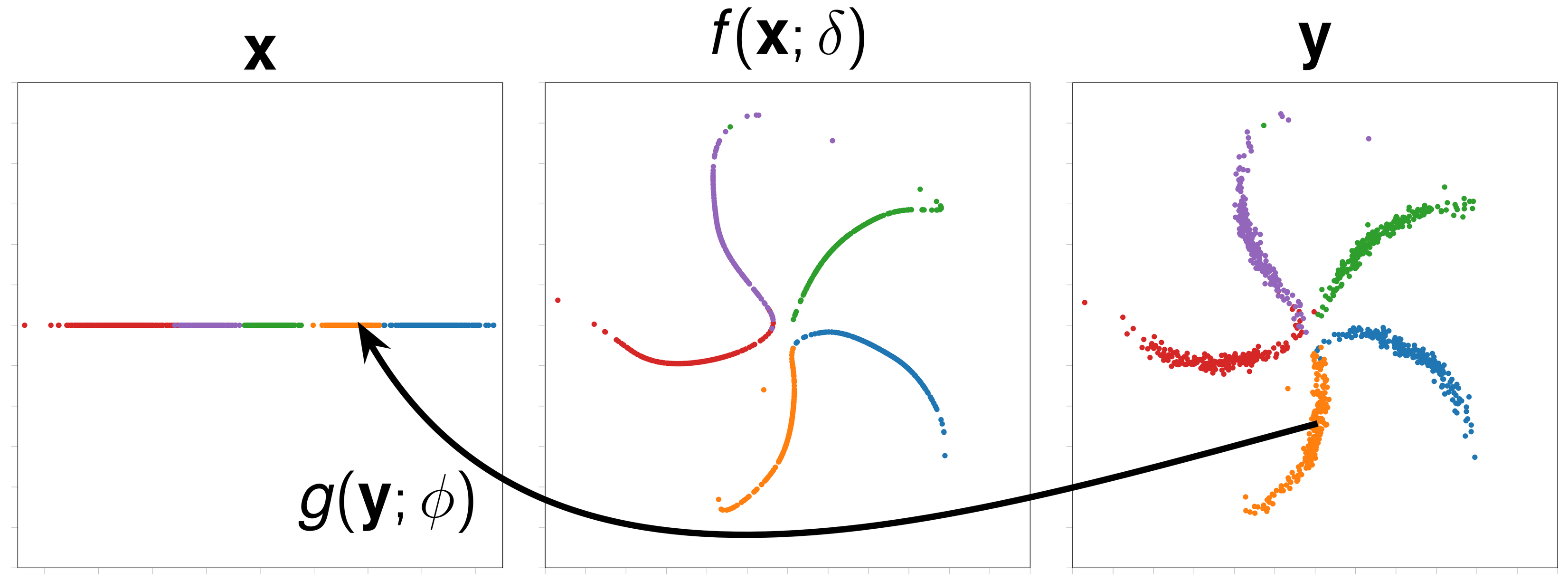
$$q(\mathbf{x}|\mathbf{y};\phi) = \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right)$$
$$\left(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \ln\boldsymbol{\sigma}_{\mathbf{x}|\mathbf{y}}\right) = g(\mathbf{y};\phi)$$



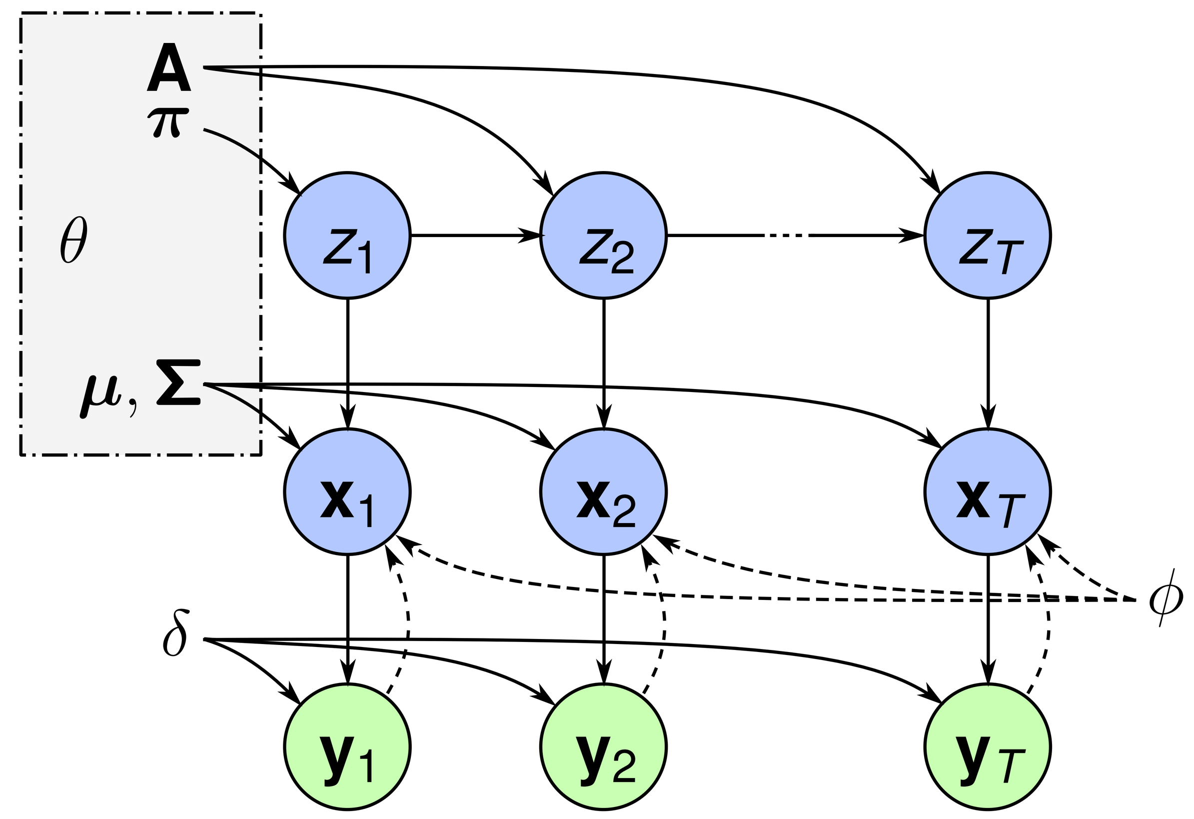


HMMVAE

- Introducing latent classes (states) z
- Class specific codes: $p(\mathbf{x}|z;\theta) = \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z\right)$



- Modeling temporal correlations by HMM:



Inference & Training

- Acoustic score: $\ln b_z(\mathbf{y}) = -H\left(q(\mathbf{x}|\mathbf{y}), p(\mathbf{x}|z)\right)$ enables forward-backward and Viterbi:

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

- Joint training of all parameters θ, δ, ϕ
- Objective: Maximization of variational lower bound

$$\mathcal{L}(\mathbf{Y};\theta,\delta,\phi) = \mathbb{E}_{q(\mathbf{X}|\mathbf{Y};\phi)}\left[\ln p(\mathbf{Y}|\mathbf{X};\delta)\right] - \mathrm{KL}\left(q(\mathbf{X},\mathbf{Z}|\mathbf{Y};\phi)||p(\mathbf{X},\mathbf{Z};\theta)\right)$$

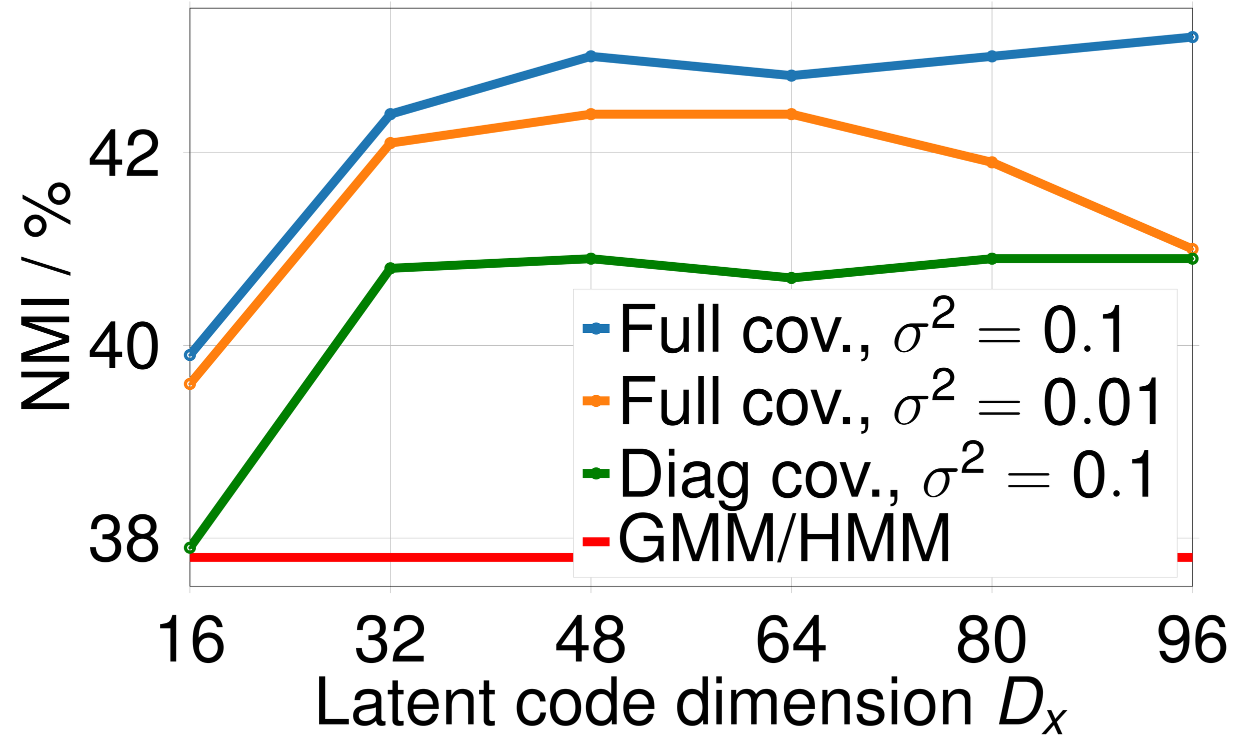
- Optimizer: Adam with stepsize α=10^{-3}

Experiments

- Goal: discovering latent acoustic units
- Database: Timit
  - 4620 train sentences (3.14h), 1680 test sentences (0.81h)
- Features: 13 element MFCCs with Δ and $\Delta\Delta$
- Initialization:
  - Using segmentations found by unsupervised GMM/HMM [2]
  - 72 AUs, each modeled by three states (left-right topology)

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

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



Conclusions

- Extended VAE by an HMM in latent code space to capture temporal correlations
- 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
- Outlook: Bayesian parameter estimation

References

[1] "Auto-Encoding Variational Bayes", D. P. Kingma and M. Welling
[2] "Variational Inference for Acoustic Unit Discovery", L. Ondel, L. Burget, and J. Cernocky

Interspeech 2017, Stockholm
This work was in part supported by Deutsche Forschungsgemeinschaft (DFG) under contract no Ha3455/12-1.