Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery

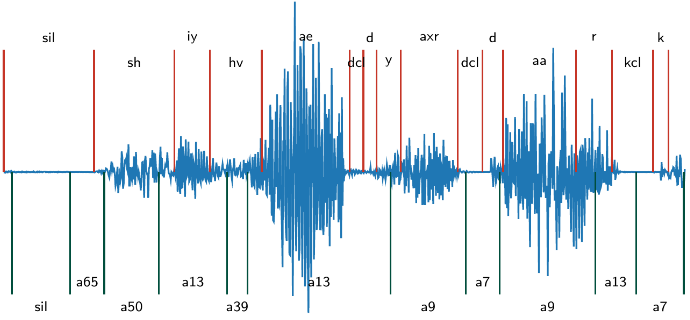
Thomas Glarner, Patrick Hanebrink, Janek Ebbers, Reinhold Haeb-Umbach

Department of Communications Engineering - Paderborn University
Prof. Dr.-Ing. Reinhold Haeb-Umbach

2018/09/05

Introduction

phones



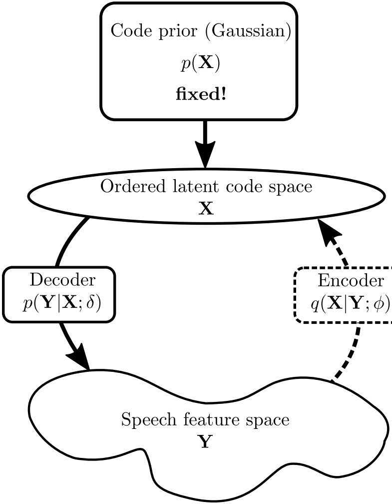
AUs

Acoustic Unit Discovery

Segment speech into resonable (phone-like) acoustic units (AUs) and simultaneously learn set of AUs

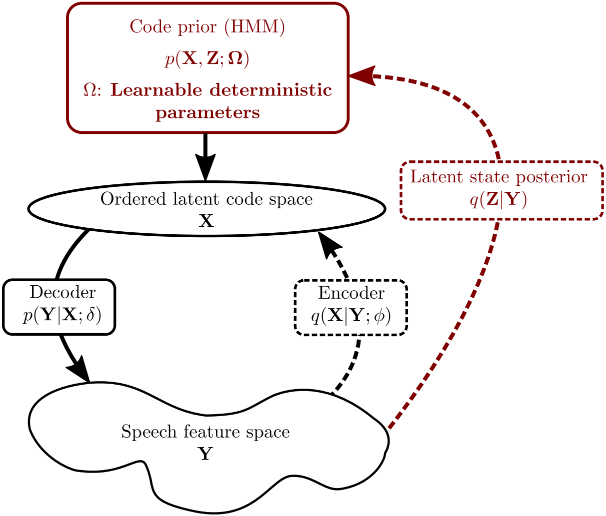
Hidden Markov Model Variational Autoencoder

Standard VAE



Hidden Markov Model Variational Autoencoder

HMM-VAE



Bayesian HMM-VAE

Drawbacks of HMM-VAE

- Number of AUs fixed up-front
- Maximum Likelihood for latent model: regularization issues

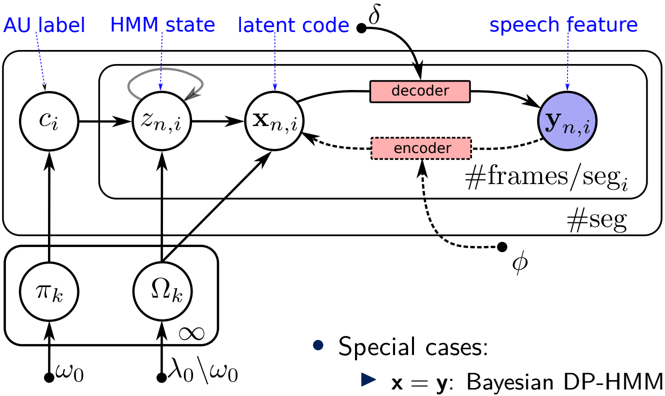
Proposed Changes

- Model should learn number of necessary AUs itself

→ Model number of HMMs/AUs with categorical distribution and Dirichlet process (DP) prior

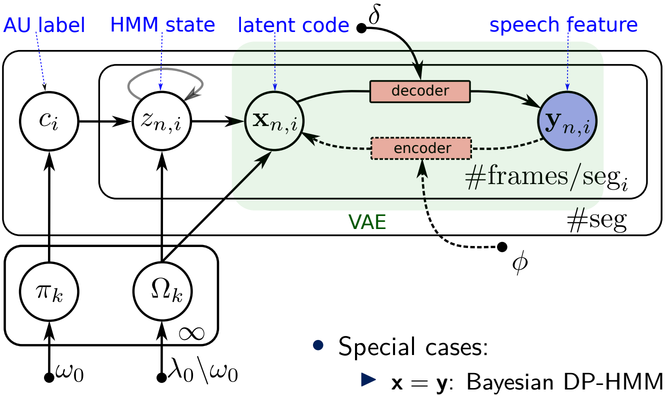
- Place conjugate priors over all latent model parameters

Graphical Model



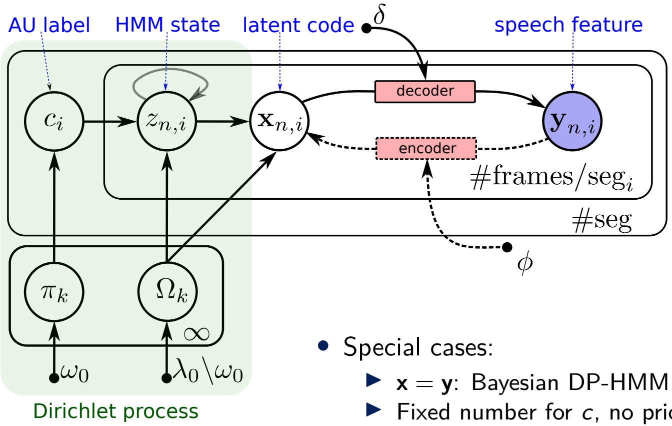
- Special cases:
  - $\mathbf{x} = \mathbf{y}$: Bayesian DP-HMM
  - Fixed number for c, no priors: HMM-VAE
- Approximate DP with finite symmetric Dirichlet distribution (truncate at U=100)

Graphical Model



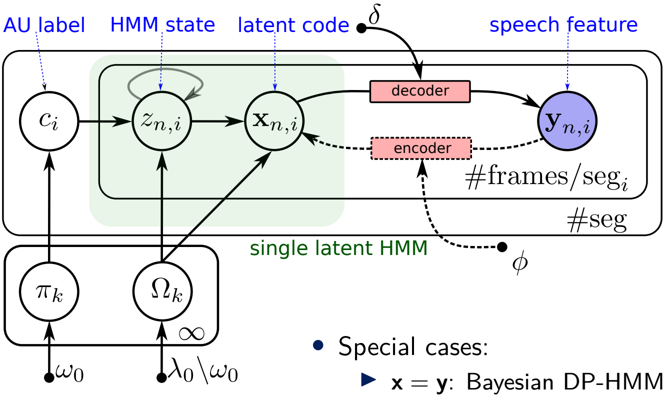
- Special cases:
  - $\mathbf{x} = \mathbf{y}$: Bayesian DP-HMM
  - Fixed number for c, no priors: HMM-VAE
- Approximate DP with finite symmetric Dirichlet distribution (truncate at U=100)

Graphical Model



- Special cases:
  - $\mathbf{x} = \mathbf{y}$: Bayesian DP-HMM
  - Fixed number for c, no priors: HMM-VAE
- Approximate DP with finite symmetric Dirichlet distribution (truncate at U=100)

Graphical Model



- Special cases:
  - $\mathbf{x} = \mathbf{y}$: Bayesian DP-HMM
  - Fixed number for c, no priors: HMM-VAE
- Approximate DP with finite symmetric Dirichlet distribution (truncate at U=100)

Training

Cost Function: Evidence Lower Bound (ELBO)

Insight: Decompose into three distinct terms:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{X};\phi)}\left[\log p(\mathbf{Y}|\mathbf{X};\delta)\right] + \mathrm{H}(q(\mathbf{X};\phi)) + \mathbb{E}_{q(\mathbf{X};\phi)}\left[\underbrace{\mathbb{E}_{q(\mathbf{Z},\mathbf{C},\Omega;\lambda)}\left[\log \frac{p(\mathbf{X},\mathbf{Z},\mathbf{C},\Omega;\lambda_0)}{q(\mathbf{Z},\mathbf{C},\Omega;\lambda)}\right]}_{\text{ELBO of Bayesian DP-HMM with “observations” } \mathbf{X}}\right].$$

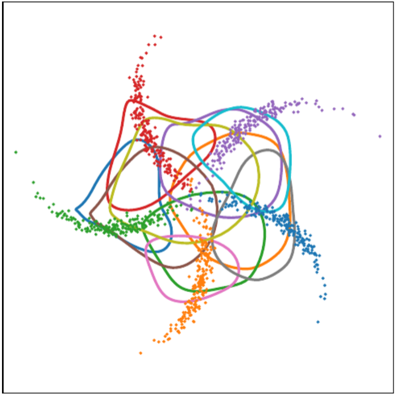
Consequences

- Decoder NN training needs first term (minibatch SGD)
- Encoder NN training needs all three terms (minibatch SGD)
- Latent model training needs third term (Stochastic Variational Inference (SVI))

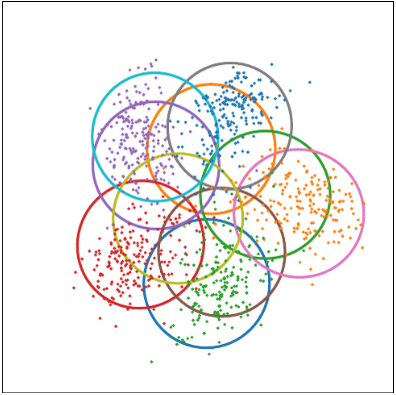
Example: Pinwheel

Synthetic pinwheel dataset, Bayesian GMM-VAE used as illustration

observation space



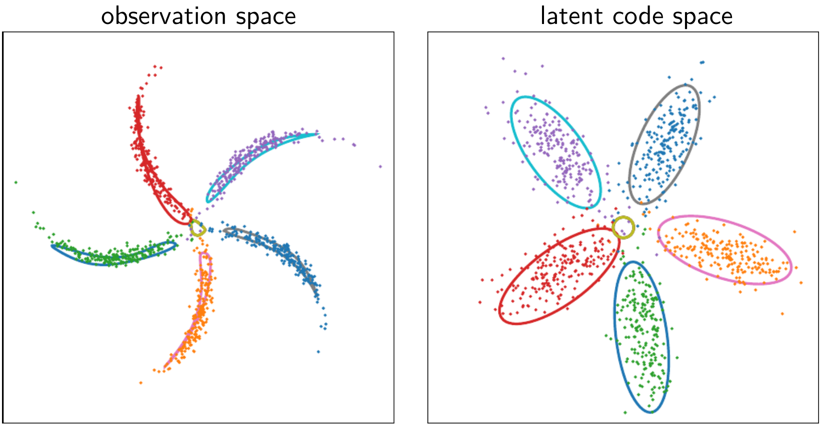
latent code space



Epoch 1

Example: Pinwheel

Synthetic pinwheel dataset, Bayesian GMM-VAE used as illustration



Epoch 500

Experimental Setup

- Datasets:
  - TIMIT
  - Xitsonga
- Measures:
  - **Normalized Mutual Information (NMI, higher is better)**: Use confusion matrix between AUs and ground truth phones, calculate mutual information and divide by ground truth phone entropy
  - **Equivalent Phone Error Rate (PER, lower is better)**: Use confusion matrix to define mapping from AU to most overlapping ground truth phone, translate AU into phone alignments, remove repititions and calculate error rate wrt. ground truth phone alignment.
- Varied parameters:
  - Emission covariance type: (**cov type**, Full more flexible than Diag)
  - SVI learning rate (**SVI lr**, matching with NN learning rate)
  - DP concentration (**DPC**, higher means fewer units are pruned)

Results on TIMIT

model/cov type	SVI lr	DPC	PER	NMI	#AU
GMM-HMM/Diag	-	-	65.42	37.84	72
HMM-VAE/Full	-	-	58.54	43.90	72
BHMMVAE/Diag	0.0010	1.000	58.74	45.08	72
	0.0010	0.100	**56.57**	**45.97**	85
	0.0010	0.010	57.31	44.58	87
	0.0100	0.010	63.12	38.81	37

Same number of AUs forced: Improved Performance

Results on TIMIT

model/cov type	SVI lr	DPC	PER	NMI	#AU
GMM-HMM/Diag	-	-	65.42	37.84	72
HMM-VAE/Full	-	-	58.54	43.90	72
BHMMVAE/Diag	0.0010	1.000	58.74	45.08	72
	0.0010	0.100	**56.57**	**45.97**	85
	0.0010	0.010	57.31	44.58	87
	0.0100	0.010	63.12	38.81	37

Best result with same learning rate and reduced concentration

Results on TIMIT

model/cov type	SVI lr	DPC	PER	NMI	#AU
GMM-HMM/Diag	-	-	65.42	37.84	72
HMM-VAE/Full	-	-	58.54	43.90	72
BHMMVAE/Diag	0.0010	1.000	58.74	45.08	72
	0.0010	0.100	**56.57**	**45.97**	85
	0.0010	0.010	57.31	44.58	87
	0.0100	0.010	63.12	38.81	37

Concentration too low: slight performance loss

Results on TIMIT

model/cov type	SVI lr	DPC	PER	NMI	#AU
GMM-HMM/Diag	-	-	65.42	37.84	72
HMM-VAE/Full	-	-	58.54	43.90	72
BHMMVAE/Diag	0.0010	1.000	58.74	45.08	72
	0.0010	0.100	**56.57**	**45.97**	85
	0.0010	0.010	57.31	44.58	87
	0.0100	0.010	63.12	38.81	37

Low concentration, learning rate too high: performance breaks down
(However: still better than GMM-HMM, considerably fewer AUs)

Results on Xitsonga

model	Cov type	SVI lr	DP C	PER	NMI	#AU
GMM-HMM	Diag	-	-	72.60	35.00	69
HMM-VAE	Full	-	-	61.90	37.60	69
BHMMVAE	Diag	0.001	1.000	62.65	37.08	69
	Full	0.001	1.000	62.64	37.08	69
	Diag	0.001	0.100	62.09	**40.06**	100
	Full	0.001	0.010	62.57	37.06	100
	Full	0.005	0.010	61.97	39.67	61

Same number of AUs forced: Only slight difference between full and diag

Results on Xitsonga

model	Cov type	SVI lr	DP C	PER	NMI	#AU
GMM-HMM	Diag	-	-	72.60	35.00	69
HMM-VAE	Full	-	-	61.90	37.60	69
BHMMVAE	Diag	0.001	1.000	62.65	37.08	69
	Full	0.001	1.000	62.64	37.08	69
	Diag	0.001	0.100	62.09	**40.06**	100
	Full	0.001	0.010	62.57	37.06	100
	Full	0.005	0.010	61.97	39.67	61

Best result (for NMI): low learning rate, but many AUs

Results on Xitsonga

model	Cov type	SVI lr	DP C	PER	NMI	#AU
GMM-HMM	Diag	-	-	72.60	35.00	69
HMM-VAE	Full	-	-	61.90	37.60	69
BHMMVAE	Diag	0.001	1.000	62.65	37.08	69
	Full	0.001	1.000	62.64	37.08	69
	Diag	0.001	0.100	62.09	**40.06**	100
	Full	0.001	0.010	62.57	37.06	100
	Full	0.005	0.010	61.97	39.67	61

Full covariance matrices: Good result possible, but well tuned learning rate needed!

Conclusions

- Bayesian priors lead to improved results
- Including a Dirichlet Process prior allows the model to autonomously infer the number of AUs
- Outcomes reasonably robust wrt. DP concentration
- SVI allows learning of probabilistic models in concert with NNs, but well matched learning rates necessary to obtain good results

Backup: Stochastic Variational Inference

SVI (Hoffmann)

- $\hat{\lambda}_n$ are natural posterior parameter values for current example
- Natural gradient for ELBO (single example): $\tilde{\nabla}_\lambda \mathcal{L} = \hat{\lambda}_n - \lambda$
- Gradient update: $\lambda_{n+1} = \lambda_n + \tau \left(\hat{\lambda}_n - \lambda_n \right) = (1 - \tau)\lambda_n + \tau \hat{\lambda}_n$
- Extend to minibatch algorithm: $\hat{\lambda}_m = \frac{N}{M_m} \sum_{n \in \mathcal{M}_m} \hat{\lambda}_n$

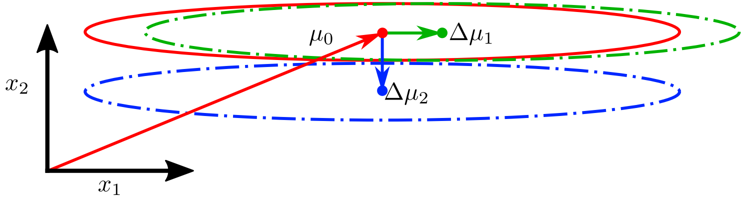
Advantage of SVI

- Batch VI leads to model training velocity mismatch
- SVI enables minibatch algorithm
- Two different learning rates to match

Backup: Natural Gradients

- Gradient of ELBO contains Hessian of normalizer:
  $\nabla_\lambda \mathcal{L} = \nabla_\lambda \nabla_\lambda^{\mathrm{T}} a(\lambda) \left(\hat{\lambda}_n - \lambda_n\right)$
- Natural gradient from information geometry:
  $\tilde{\nabla}_\lambda \mathcal{L} = \mathrm{I}(\lambda)^{-1} \nabla_\lambda \mathcal{L}$
  (Works better, but requires inverse of fisher information matrix)
- For exponential family: $\mathrm{I}(\lambda) = \nabla_\lambda \nabla_\lambda^{\mathrm{T}} a(\lambda)$

$\Rightarrow$ Natural gradient simplifies gradient calculation for ELBO!



Backup: Initialization options

pre-train

Pseudo-supervised pretraining (20 epochs) with randomly generated alignment (fixed length) as label sequence for each utterance

cluster

Initialize latent space with standard VAE and perform k-means clustering ($k = 3U$) on latent space to initialize state distributions