A Benchmark of Dynamical Variational Autoencoders applied to Speech Spectrogram Modeling

Xiaoyu Bie[1], Laurent Girin[2], Simon Leglaive[3],
Thomas Hueber[2] and Xavier Alameda-Pineda[1]

[1] Inria, Univ. Grenoble Alpes, CNRS, LJK, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, 38000 Grenoble, France
[3] CentraleSupélec, IETR, 35576 Cesson-Sévigné, France

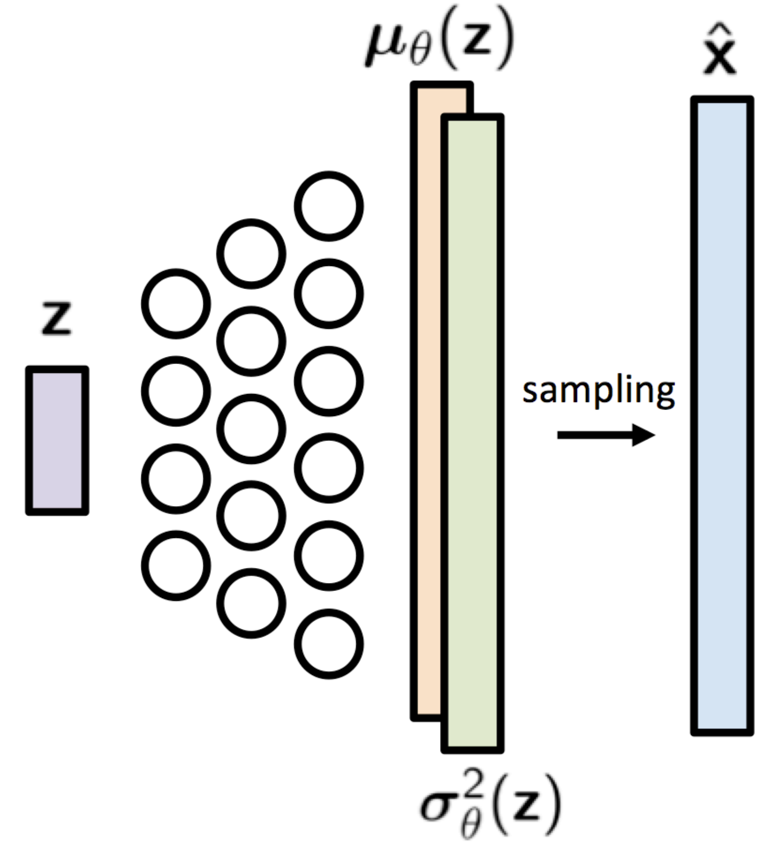
This research was supported by ANR-3IA MIAI, ANR-JCJC ML3RI and H2020 SPRING

Part 1:

From VAE to Dynamical VAE

Variational Autoencoder (VAE)

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}) d\mathbf{z}$$



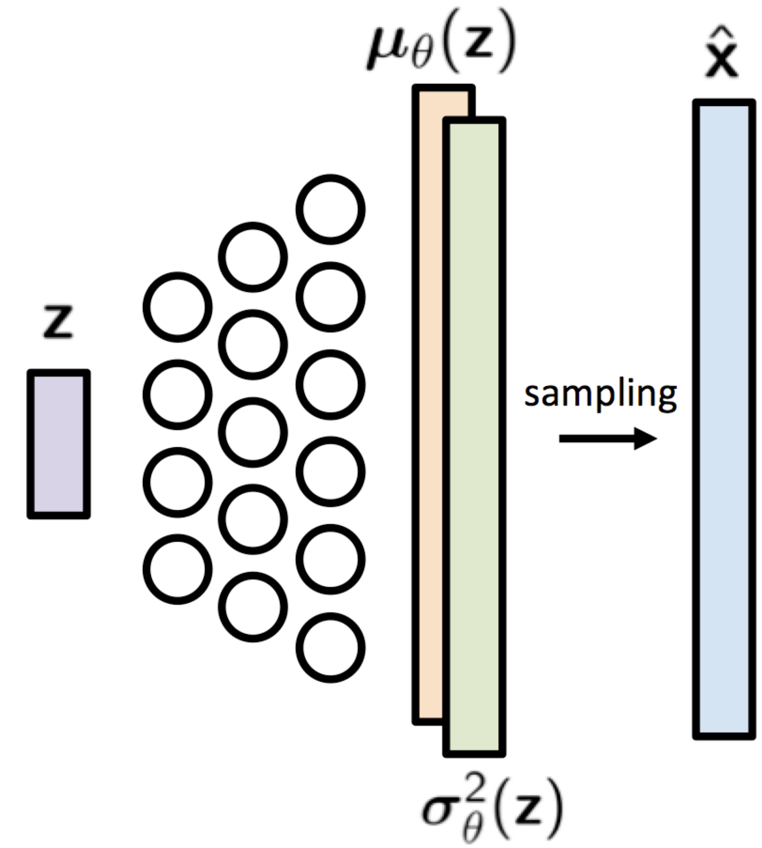
- VAE is a deep generative model, $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ (decoder) is defined via a DNN (e.g. MLP)
- For example, $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ can be a Gaussian with mean and variance being the output of the DNN with input $\mathbf{z}$
- Directly computing $p_{\boldsymbol{\theta}}(\mathbf{x})$ for parameter estimation is intractable

Variational Autoencoder (VAE)

$$\ln p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathscr{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\varphi}) + D_{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}\,|\,\mathbf{x})\right]$$

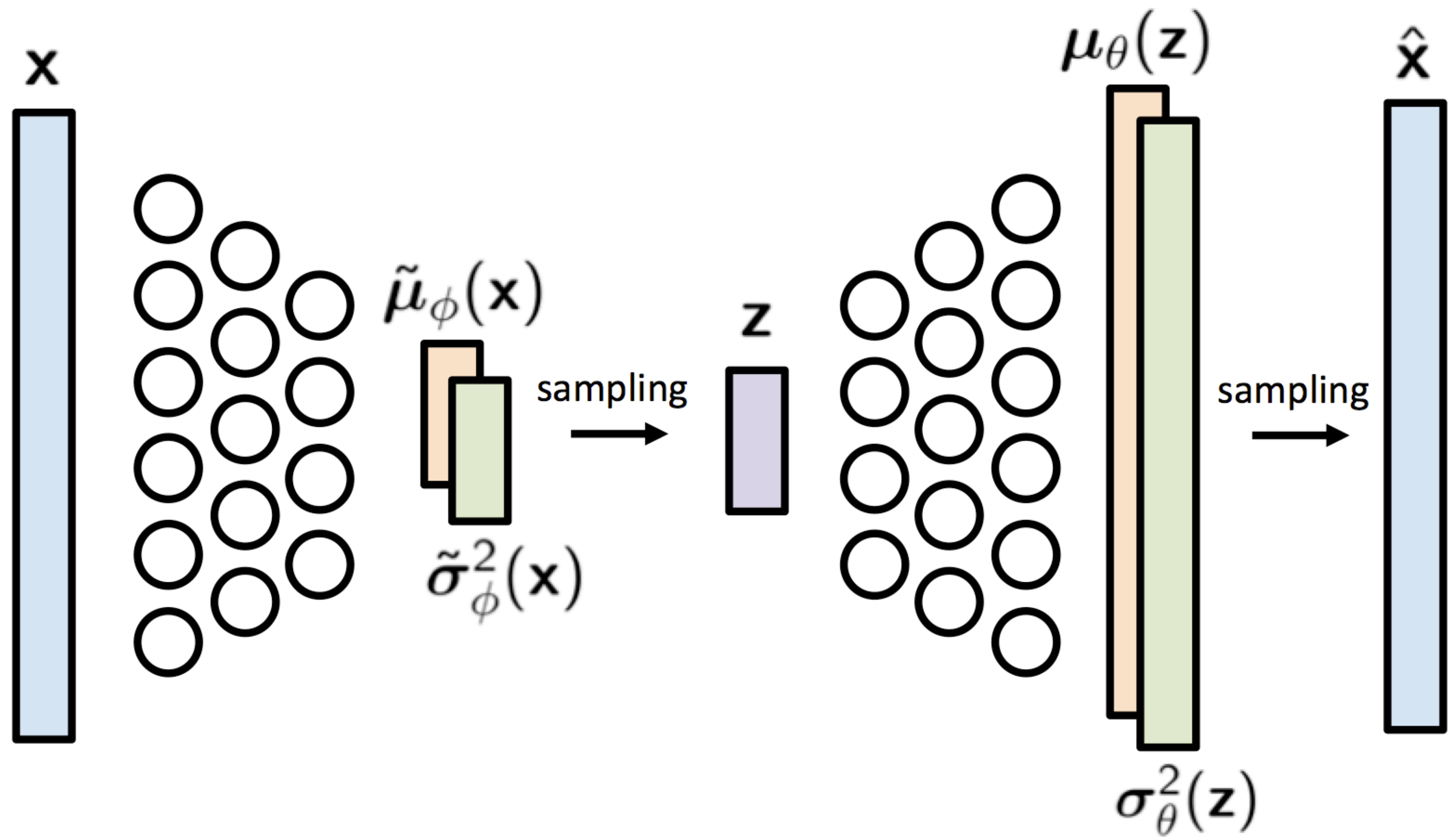
where

$$\mathscr{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\varphi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})\right] - D_{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}) \parallel p(\mathbf{z})\right]$$



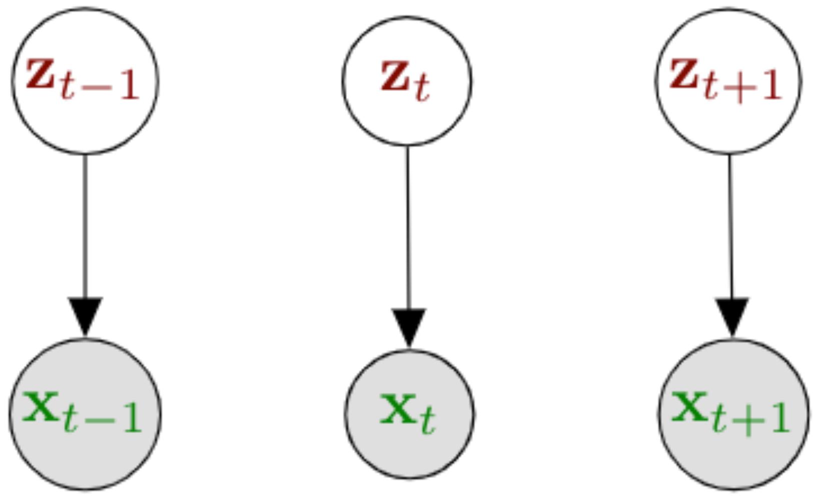
- VAE is a deep generative model, $p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})$ (decoder) is defined via a DNN (e.g. MLP)
- For example, $p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})$ can be a Gaussian with mean and variance being the output of the DNN with input $\mathbf{z}$
- Directly computing $p_{\boldsymbol{\theta}}(\mathbf{x})$ for parameter estimation is intractable
- $\mathscr{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\varphi})$ is the evidence lower bound (ELBO), where $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})$ is the variational approximate posterior distribution

Variational Autoencoder (VAE)



- VAE is a deep generative model, $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ (decoder) is defined via a DNN (e.g. MLP)
- For example, $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ can be a Gaussian, with mean and variance being the output of the DNN with input $\mathbf{z}$
- Directly computing $p_{\boldsymbol{\theta}}(\mathbf{x})$ for parameter estimation is intractable
- $\mathscr{L}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\varphi})$ is the evidence lower bound (ELBO), where $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ is the variational approximate posterior distribution
- A VAE model is trained by cascading the encoder and decoder and maximizing the ELBO w.r.t. both encoder and decoder parameters

From VAE to Dynamical VAE (DVAE)

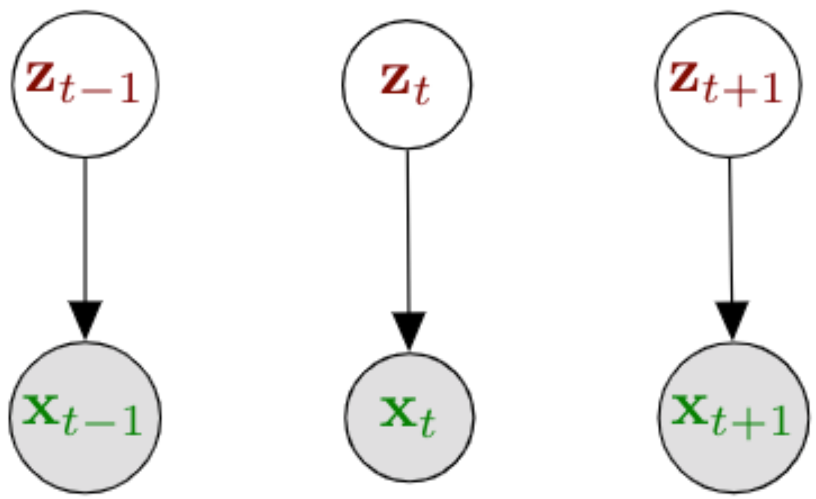


VAE

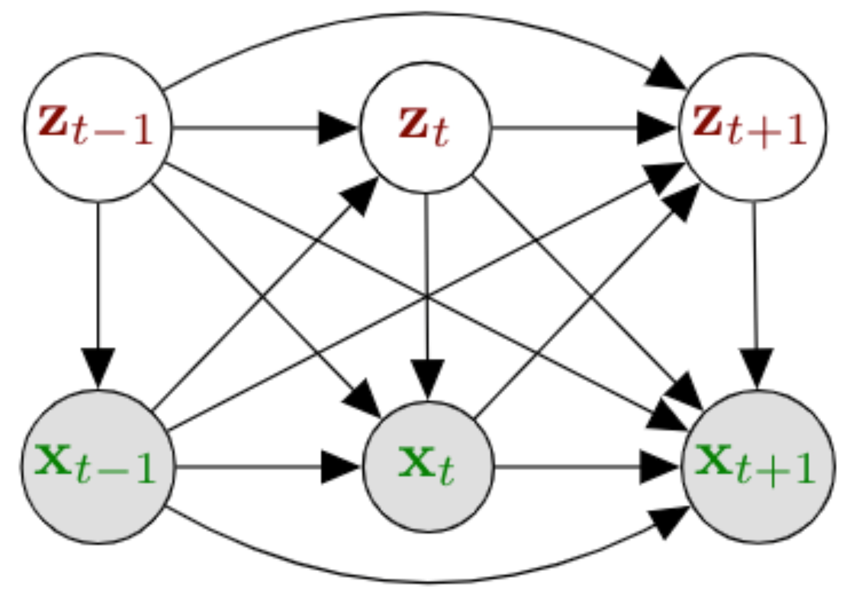
$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} \int p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{z}_t) p_{\boldsymbol{\theta}}(\mathbf{z}_t) d\mathbf{z}_t$$

- Major limitation of VAE: All vector pairs $(\mathbf{x}_t, \mathbf{z}_t)$ are assumed independent
- Problem: There is correlation between frames for sequential data, VAE is too simple

From VAE to Dynamical VAE (DVAE)



VAE



DVAE

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} \int p_{\boldsymbol{\theta}}(\mathbf{x}_t \,|\, \mathbf{z}_t) p_{\boldsymbol{\theta}}(\mathbf{z}_t) d\mathbf{z}_t$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) d\mathbf{z}_{1:T}$$

- Major limitation of VAE: all vector pairs $(\mathbf{x}_t, \mathbf{z}_t)$ are assumed independent
- Problem: There is correlation between frames for sequential data, VAE is too simple
- DVAE is the generalization of VAE to correlated sequential data
- DVAE is a family of models obtained with different simplifications of the dependencies
- DVAE are trained using the same methodology as for the VAE

Part 2:

DVAE family

DVAE family

Unified generative equation for a DVAE model:

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) p_{\boldsymbol{\theta}_z}(\mathbf{z}_t \,|\, \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$$

Simplifications of the dependencies for different DVAE models

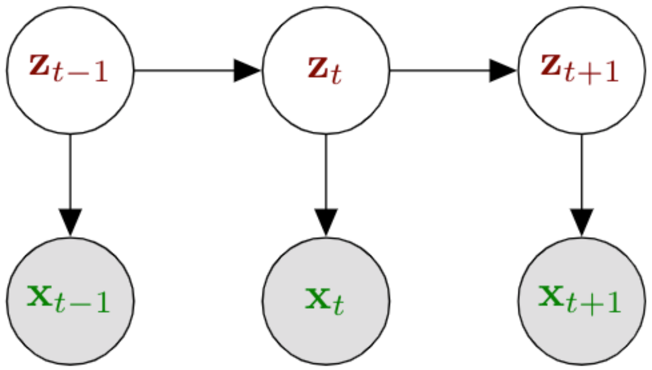
		$p_\theta(\mathbf{z}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$	$p_\theta(\mathbf{x}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$
VAE*	[Kingma and Welling, 2014, Rezende et al., 2014]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{z}_t)$
RVAE*	[Leglaive et al., 2020]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{z}_{1:t})$
STORN	[Bayer and Osendorfer, 2014]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$
DKF*	[Krishnan et al., 2015, Krishnan et al., 2017]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t \vert \mathbf{z}_{t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{z}_t)$
DSAE	[Li and Mandt, 2018]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t \vert \mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{z}_t, \mathbf{v})$
VRNN	[Chung et al., 2015, Goyal et al., 2017]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$
SRNN*	[Fraccaro et al., 2016]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t \vert \mathbf{x}_{1:t-1}, \mathbf{z}_t)$

Two examples of DVAE

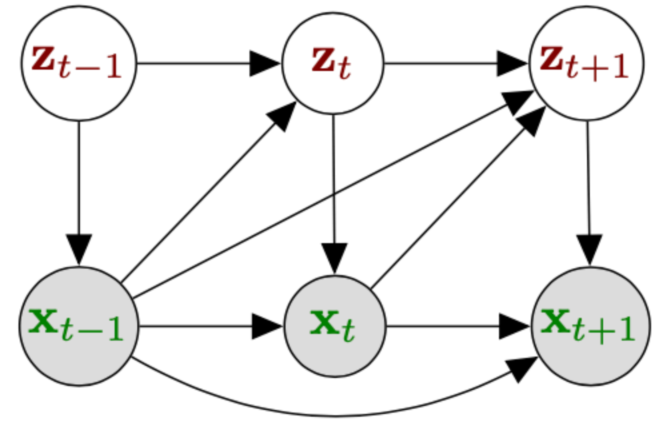
General formulation: $$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t | \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) p_{\boldsymbol{\theta}_z}(\mathbf{z}_t | \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$$

Two examples of DVAE

General formulation: $p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\theta_{\mathbf{x}}}(\mathbf{x}_t \mid \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t \mid \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$



Generation



$$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\theta_{\mathbf{x}}}(\mathbf{x}_t \mid \mathbf{z}_t) p_{\theta_z}(\mathbf{z}_t \mid \mathbf{z}_{t-1})$$

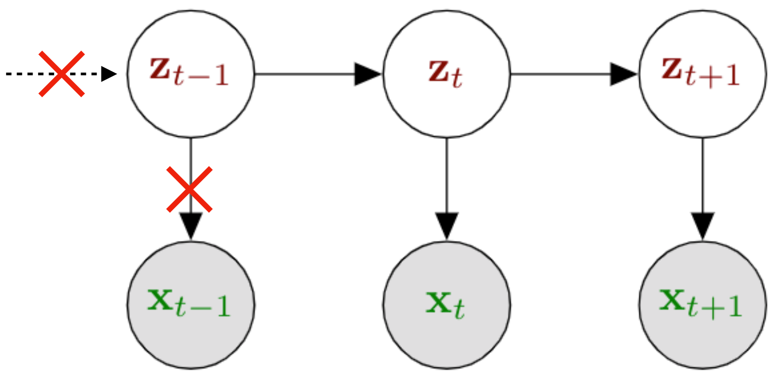
A simple SSM-like generative model

$$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\theta_{\mathbf{x}}}(\mathbf{x}_t \mid \mathbf{z}_t, \mathbf{x}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{x}_{1:t-1})$$

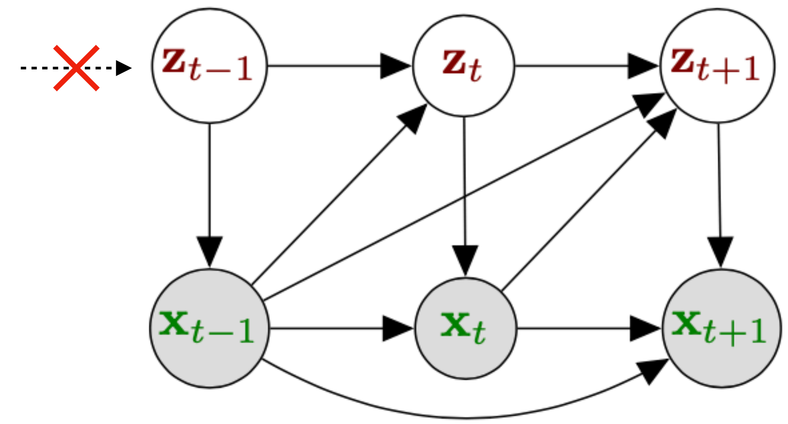
Add previous observation $\mathbf{x}_{1:t-1}$

Two examples of DVAE

General formulation: $p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) p_{\boldsymbol{\theta}_z}(\mathbf{z}_t \,|\, \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$



Generation



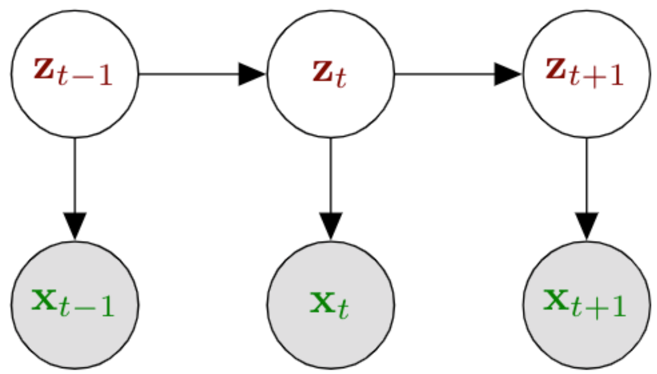
$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \approx \prod_{t=1}^{T} p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_t) p_{\boldsymbol{\theta}_z}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1})$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \approx \prod_{t=1}^{T} p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_t, \mathbf{x}_{1:t-1}) p_{\boldsymbol{\theta}_z}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{1:t-1})$$

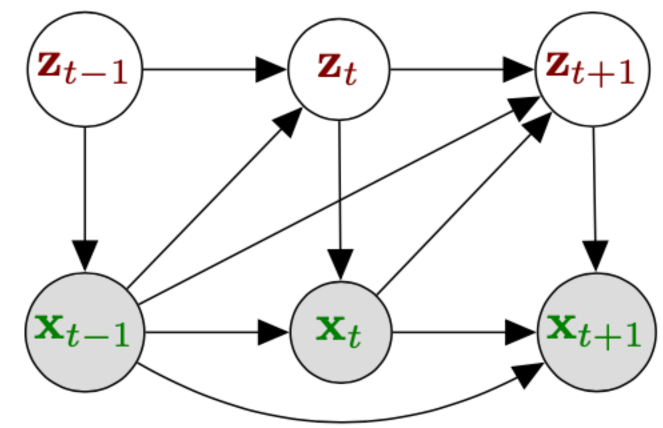
When we come to the posterior, we can apply D-separation to identify the dependencies [Bishop, 2006]

Two examples of DVAE

General formulation: $p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\theta_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t \,|\, \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$

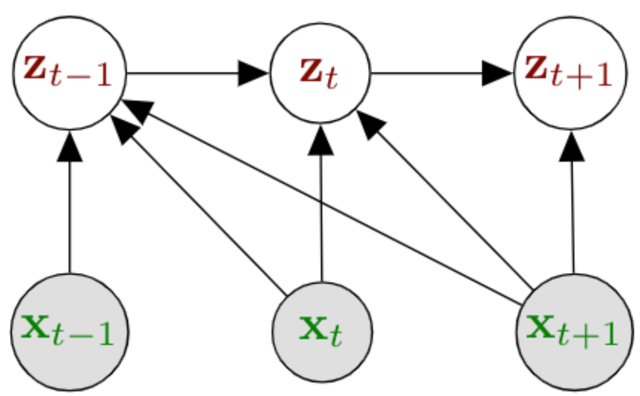


Generation

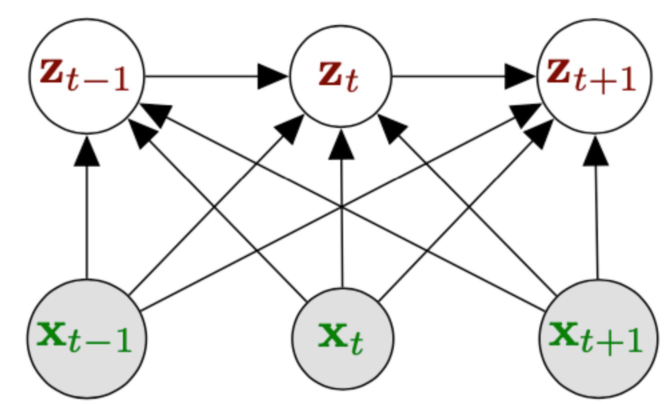


$$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \approx \prod_{t=1}^{T} p_{\theta_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_t) p_{\theta_z}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1})$$

$$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \approx \prod_{t=1}^{T} p_{\theta_{\mathbf{x}}}(\mathbf{x}_t \,|\, \mathbf{z}_t, \mathbf{x}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{1:t-1})$$



Inference



$$q_{\phi}(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_{\phi}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{t:T})$$

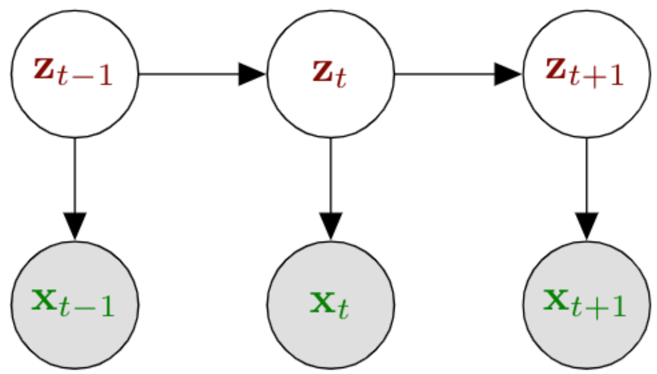
$$q_{\phi}(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_{\phi}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{1:T})$$

The inference model respects the structure of the exact posterior distribution

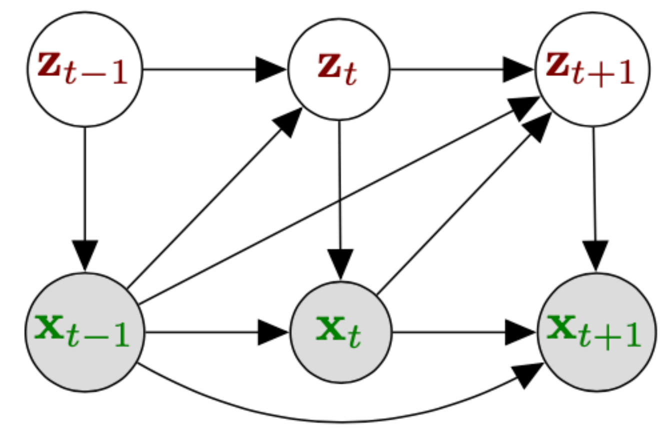
Two examples of DVAE

General formulation: $$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\theta_\mathbf{x}}(\mathbf{x}_t \,|\, \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t \,|\, \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$$

Generation

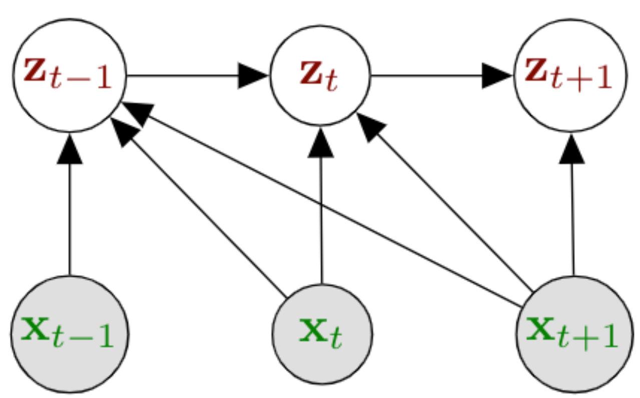


$$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \approx \prod_{t=1}^{T} p_{\theta_\mathbf{x}}(\mathbf{x}_t \,|\, \mathbf{z}_t) p_{\theta_z}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1})$$



$$p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \approx \prod_{t=1}^{T} p_{\theta_\mathbf{x}}(\mathbf{x}_t \,|\, \mathbf{z}_t, \mathbf{x}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{1:t-1})$$

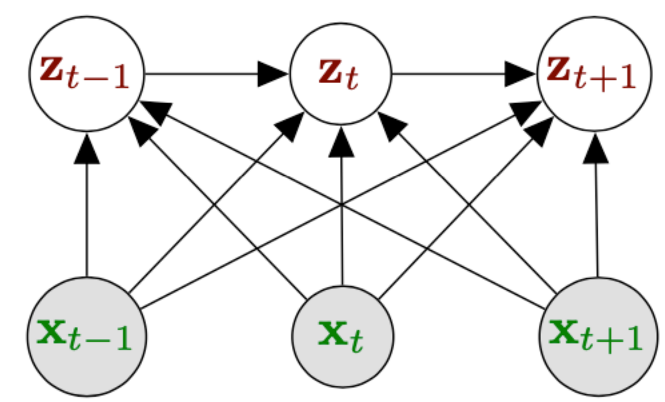
Inference



$$q_{\phi}(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_{\phi}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{t:T})$$

DKF (Krishnan et al., 2015, 2017)

Non-autoregressive DVAE



$$q_{\phi}(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_{\phi}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{x}_{1:T})$$

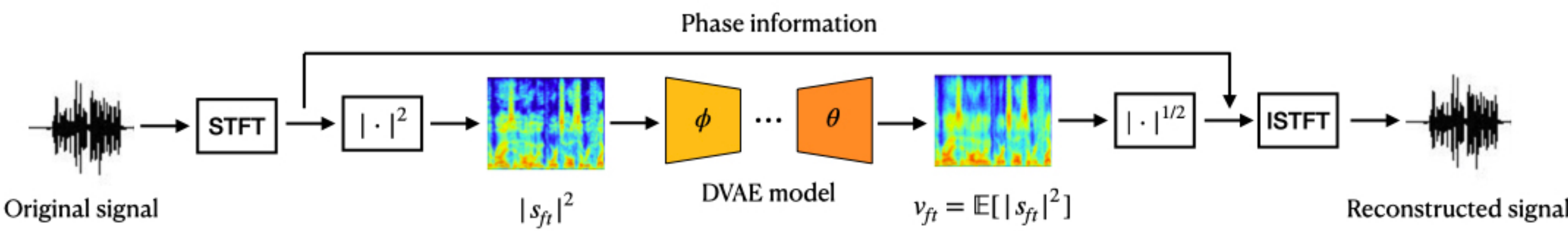
SRNN (Fraccaro et al., 2016)

Autoregressive DVAE

Part 3:

Application to speech spectrogram modeling

Analysis-resynthesis of speech signals



- Dataset: WSJ0 subsets (*si_tr_s*, *si_dt_05* and *si_et_05, different speakers*)
- Time-domain 16 kHz signals are normalized by absolute maximum value
- STFT with a 32ms sine window and 16ms hop length
- Crop the magnitude spectrogram into 150-frame sequences during training
- In summary
  - 9h for training (*si_tr_s*)
  - 1.5h for validation (*si_dt_05*)
  - 1.5h for evaluation (*si_et_05, no cropping*)

Results

	VAE	DKF	STORN	VRNN	SRNN	RVAE	DSAE
Autoregressive			✓	✓	✓		
True Posterior		✓			✓	✓	
Dynamical model on $\mathbf{z}_t$		✓		✓	✓		✓
RMSE ($\times 10^{-2}$)	5.10	3.44	3.38	2.67	**2.48**	4.99	4.69
PESQ	2.05	3.30	3.05	3.60	**3.64**	2.27	2.32
STOI	0.86	0.94	0.93	0.96	**0.97**	0.89	0.90

- All DVAEs outperform the vanilla VAE

Results

	VAE	DKF	STORN	VRNN	SRNN	RVAE	DSAE
Autoregressive			✓	✓	✓		
True Posterior		✓			✓	✓	
Dynamical model on $\mathbf{z}_t$		✓		✓	✓		✓
RMSE ($\times 10^{-2}$)	5.10	3.44	3.38	2.67	**2.48**	4.99	4.69
PESQ	2.05	3.30	3.05	3.60	**3.64**	2.27	2.32
STOI	0.86	0.94	0.93	0.96	**0.97**	0.89	0.90

- All DVAEs outperform the vanilla VAE
- Autoregressive models are powerful in speech analysis-resynthesis

Results

	VAE	DKF	STORN	VRNN	SRNN	RVAE	DSAE
Autoregressive			✓	✓	✓		
True Posterior		✓			✓	✓	
Dynamical model on $\mathbf{z}_t$		✓		✓	✓		✓
RMSE ($\times 10^{-2}$)	5.10	3.44	3.38	2.67	**2.48**	4.99	4.69
PESQ	2.05	3.30	3.05	3.60	**3.64**	2.27	2.32
STOI	0.86	0.94	0.93	0.96	**0.97**	0.89	0.90

- All DVAEs outperform the vanilla VAE
- Autoregressive models are powerful in speech analysis-resynthesis
- It is rewarding to respect the structure of the exact posterior distribution when designing the inference model

Results

	VAE	DKF	STORN	VRNN	SRNN	RVAE	DSAE
Autoregressive			✓	✓	✓		
True Posterior		✓			✓	✓	
Dynamical model on $\mathbf{z}_t$		✓		✓	✓		✓
RMSE ($\times 10^{-2}$)	5.10	3.44	3.38	2.67	**2.48**	4.99	4.69
PESQ	2.05	3.30	3.05	3.60	**3.64**	2.27	2.32
STOI	0.86	0.94	0.93	0.96	**0.97**	0.89	0.90

- All DVAEs outperform the vanilla VAE
- Autoregressive models are powerful in speech analysis-resynthesis
- It is rewarding to respect the structure of the exact posterior distribution when designing the inference model
- It is better to apply a dynamical model on $\mathbf{z}_t$, not simply assume that it is i.i.d

Results

	VAE	DKF	STORN	VRNN	SRNN	RVAE	DSAE
Autoregressive			✓	✓	✓		
True Posterior		✓			✓	✓	
Dynamical model on $\mathbf{z}_t$		✓		✓	✓		✓
RMSE ($\times 10^{-2}$)	5.10	3.44	3.38	2.67	**2.48**	4.99	4.69
PESQ	2.05	3.30	3.05	3.60	**3.64**	2.27	2.32
STOI	0.86	0.94	0.93	0.96	**0.97**	0.89	0.90

- All DVAEs outperform the vanilla VAE
- Autoregressive models are powerful in speech analysis-resynthesis
- It is rewarding to respect the structure of the exact posterior distribution when designing the inference model
- It is better to apply a dynamical model on $\mathbf{z}_t$, not simply assume that it is i.i.d
- SRNN performs the best because it features all three properties

Conclusion

- DVAE family, great potential to model speech signals!
- Code in PyTorch is available at https://github.com/XiaoyuBIE1994/DVAE-speech
- Important considerations when designing a new DVAE model:
  - Autoregressive or non-autoregressive
  - Whether the inference model respects the structure of the exact posterior distribution
  - Whether apply a dynamical model on the latent variable z_t
- More discussion for DVAE family: *Girin L, Leglaive S, Bie X, et al. Dynamical variational autoencoders: A comprehensive review. arXiv preprint arXiv:2008.12595, 2020.*
- Application of DVAE models in unsupervised speech enhancement: *Bie X, Leglaive S, Alameda-Pineda X, et al. Unsupervised Speech Enhancement using Dynamical Variational Auto-Encoders. arXiv preprint arXiv:2106.12271, 2021.*