ICML 2019

FloWaveNet: A Generative Flow for Raw Audio

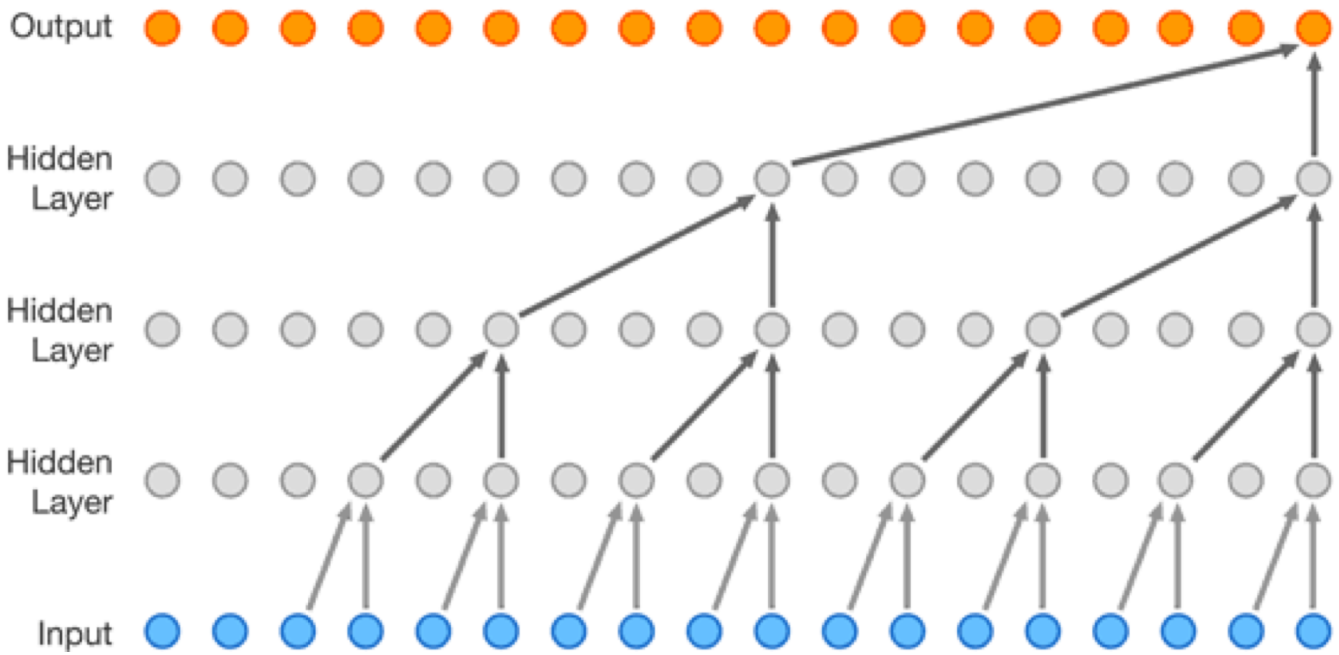
Sungwon Kim[1], Sang-gil Lee[1], Jongyoon Song[1], Jaehyeon Kim[2], Sungron Yoon[1,3]

[1]Seoul National University, [2]Kakao Corporation,
[3]ASRI, INMC, Institute of Engineering Research, Seoul National University

DSAIL Data Science & AI Laboratory kakao

Poster 6/12 6:30 PM @Pacific Ballroom #2

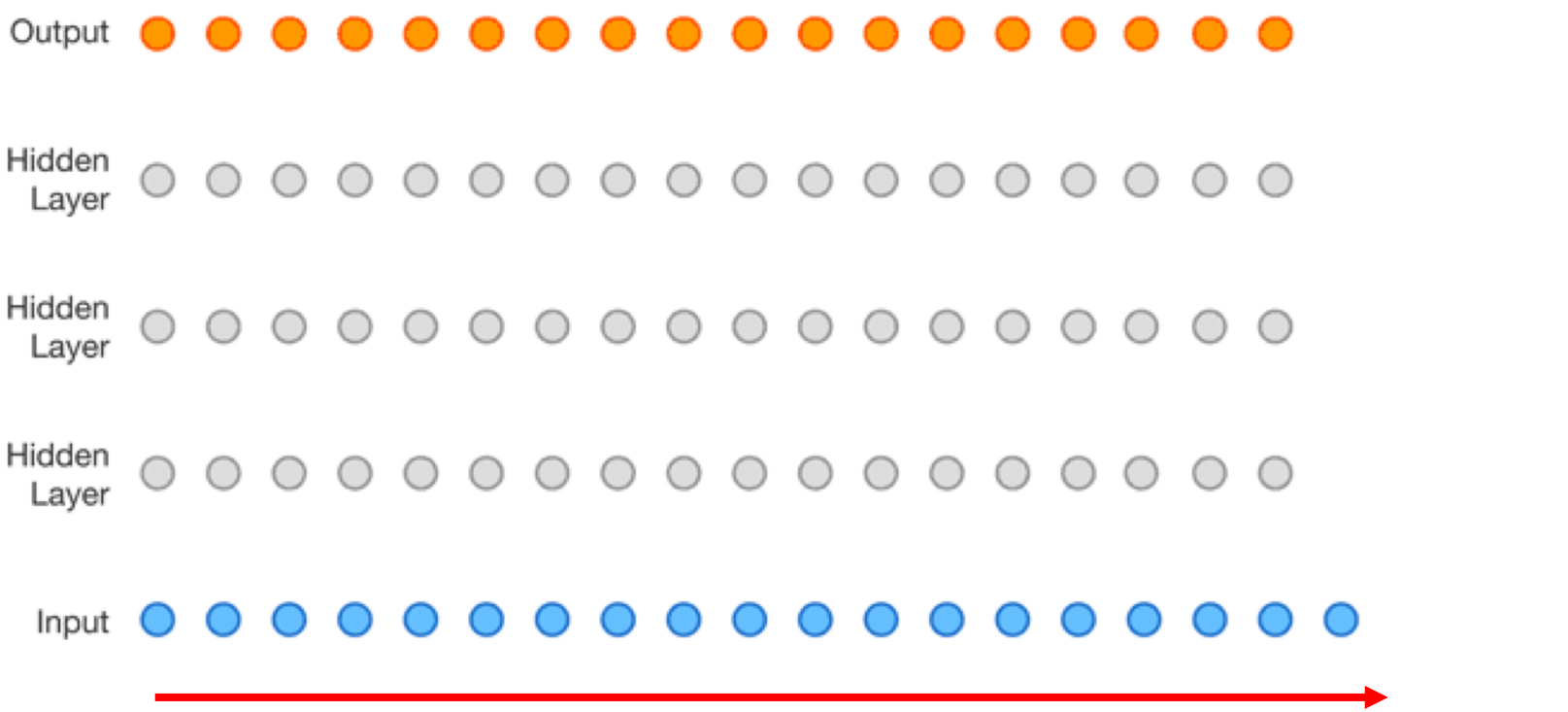
WaveNet



$$\log p_X(x_{1:T}) = \sum_{t=1}^{T} \log p_X(x_t | x_{<t})$$

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

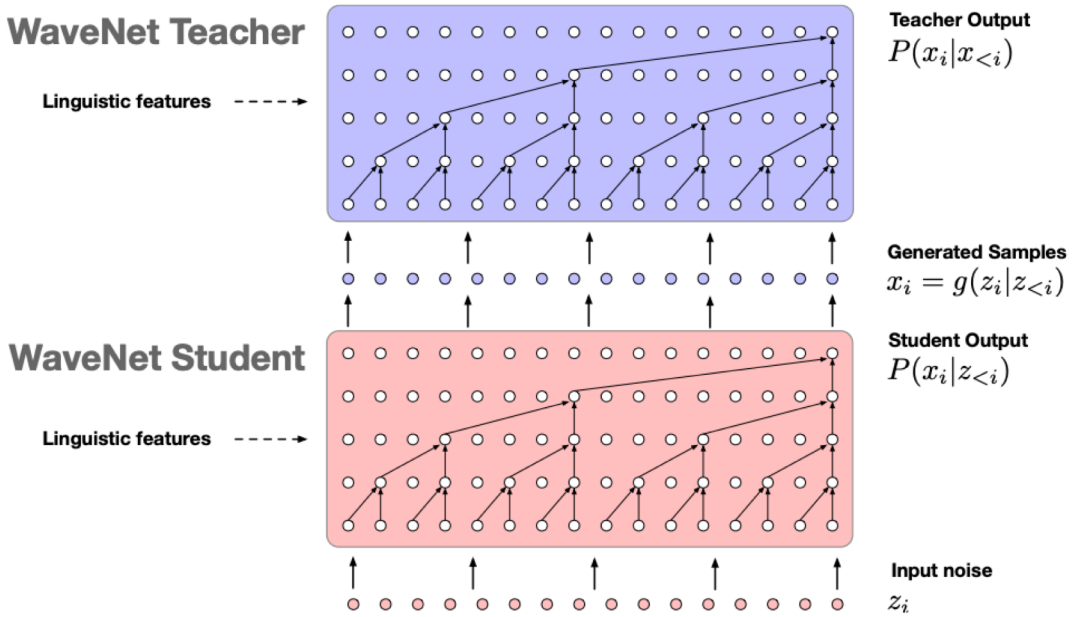
WaveNet



$$\log p_X(x_{1:T}) = \sum_{t=1}^{T} \log p_X(x_t | x_{<t})$$

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Previous parallel speech synthesis models



Pre-trained WaveNet

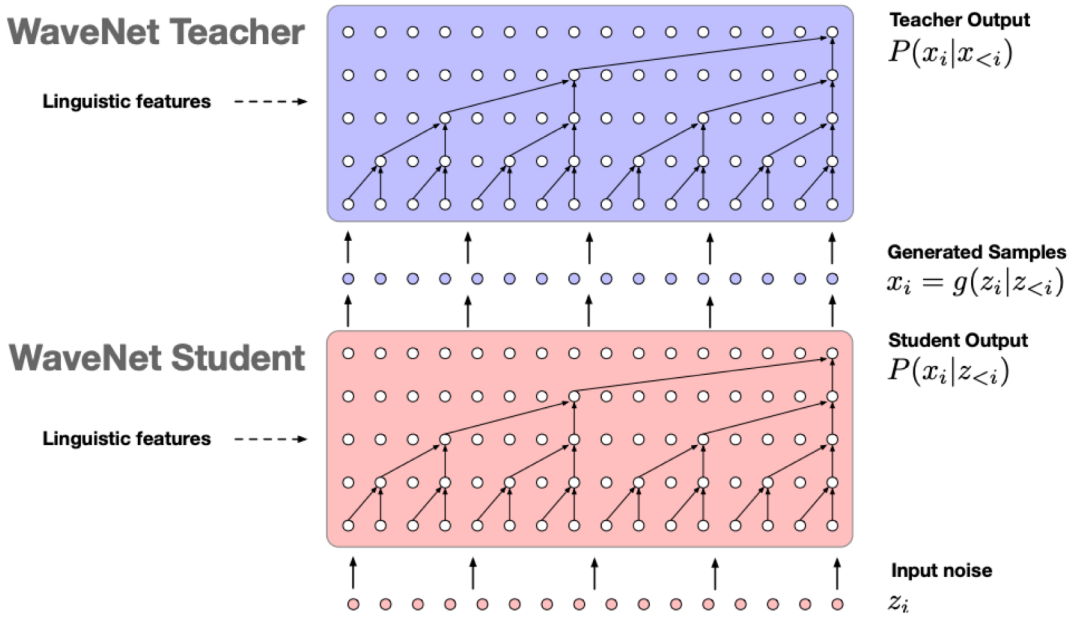
Inverse Autoregressive Flows (IAFs)

Probability Density Distillation

$$KL\left(P_S(x)||P_T(x)\right)$$

Oord, Aaron, et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis." *International Conference on Machine Learning*. 2018.

Previous parallel speech synthesis models



Pre-trained WaveNet

Parallel sampling

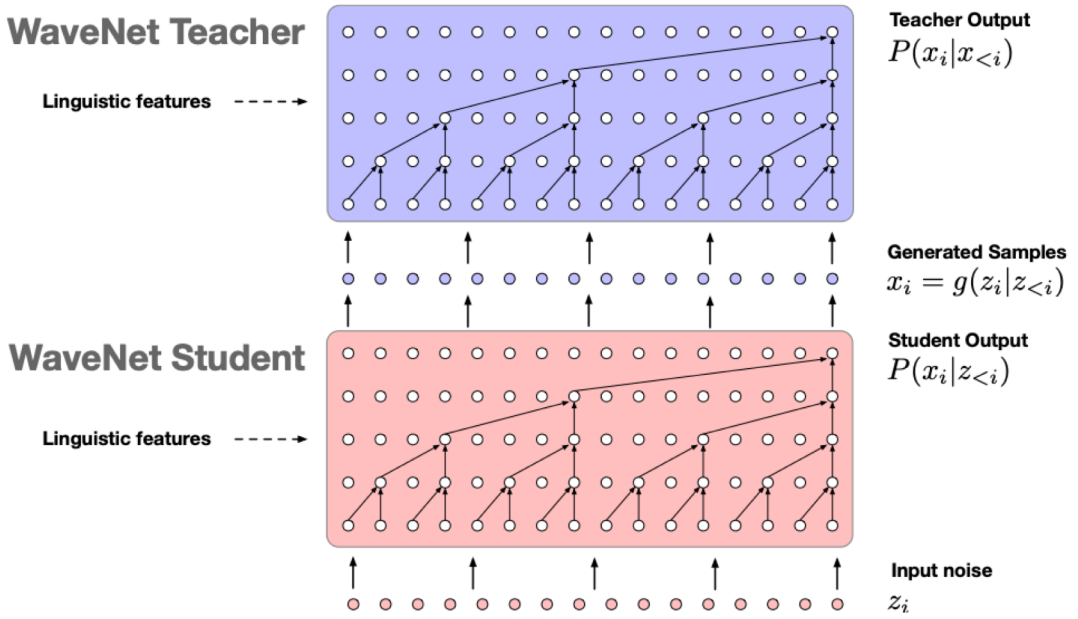
Inverse Autoregressive Flows (IAFs)

Probability Density Distillation

$$KL\left(P_S(x)||P_T(x)\right)$$

Oord, Aaron, et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis." *International Conference on Machine Learning*. 2018.

Previous parallel speech synthesis models



Pre-trained WaveNet

Parallel sampling

Inverse Autoregressive Flows (IAFs)

Probability Density Distillation

$$KL\left(P_S(x)||P_T(x)\right)$$

+

Power Loss
Perceptual Loss
Contrastive Loss
Frame Loss

Oord, Aaron, et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis." *International Conference on Machine Learning*. 2018.

Our Objectives

- Simplify the training procedure for parallel sampling
- Maintain the quality of speech samples

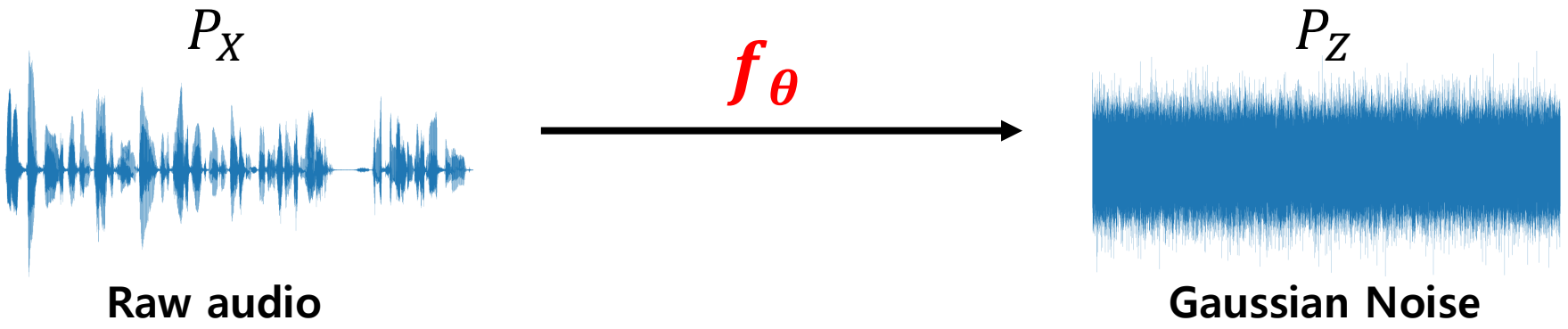
Our Objectives

- Simplify the training procedure for parallel sampling

- Maintain the quality of speech samples

Flow-based generative models for raw audio!

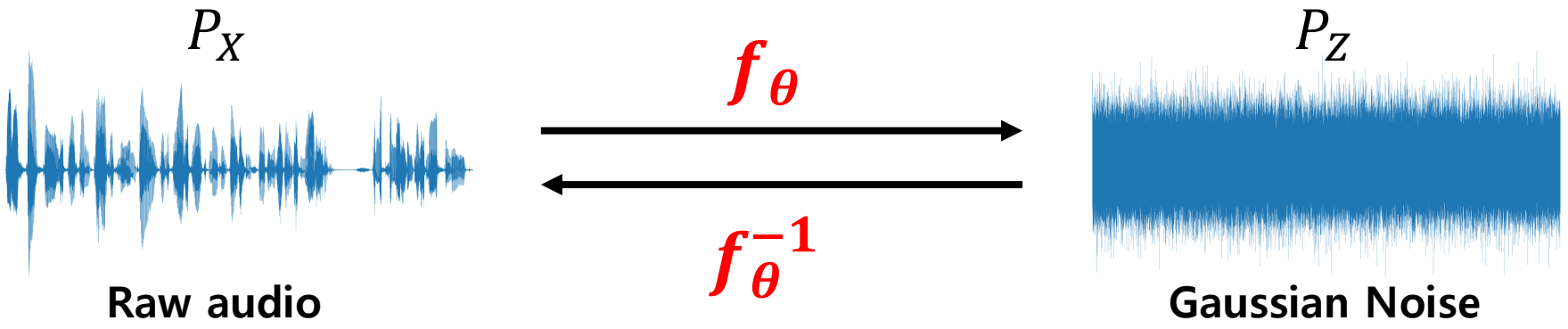
FloWaveNet



Training phase

$$\log p_X(x_{1:T}) = \log p_Z\left(f_\theta(x_{1:T})\right) + \log\det\left(\frac{\partial f_\theta(x)}{\partial x}\right)$$

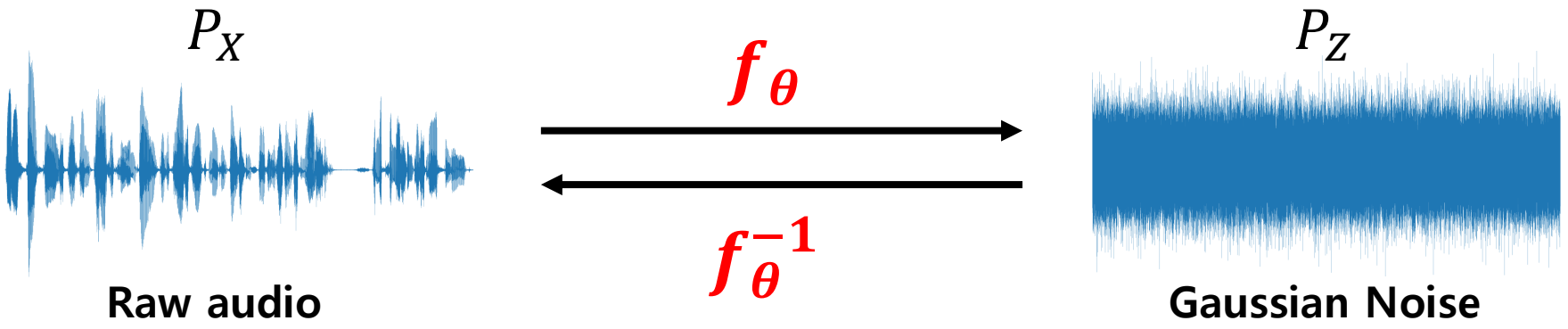
FloWaveNet



Training phase $$\log p_X(x_{1:T}) = \log p_Z\left(f_\theta(x_{1:T})\right) + \log\det\left(\frac{\partial f_\theta(x)}{\partial x}\right)$$

Sampling phase $$z = z_{1:T} \sim P_Z(z) = N(O, I), \qquad \hat{x} = f_\theta^{-1}(z)$$

FloWaveNet

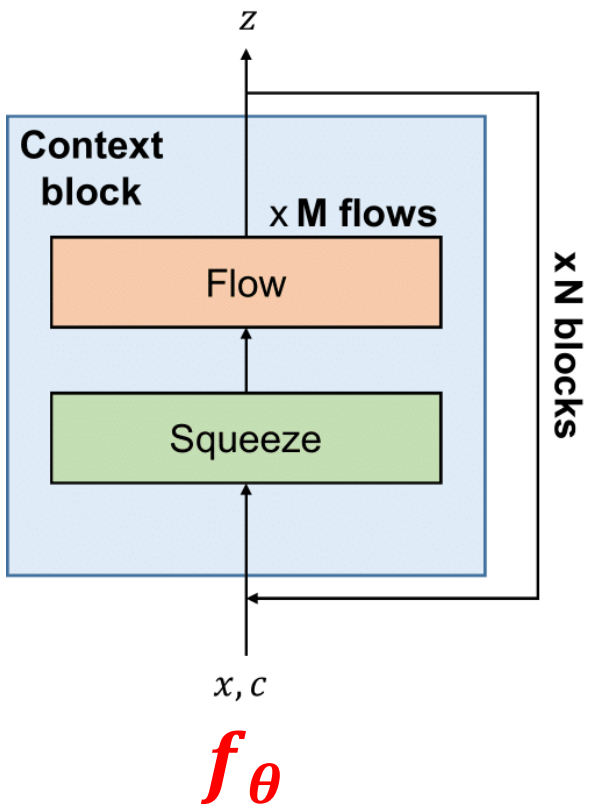


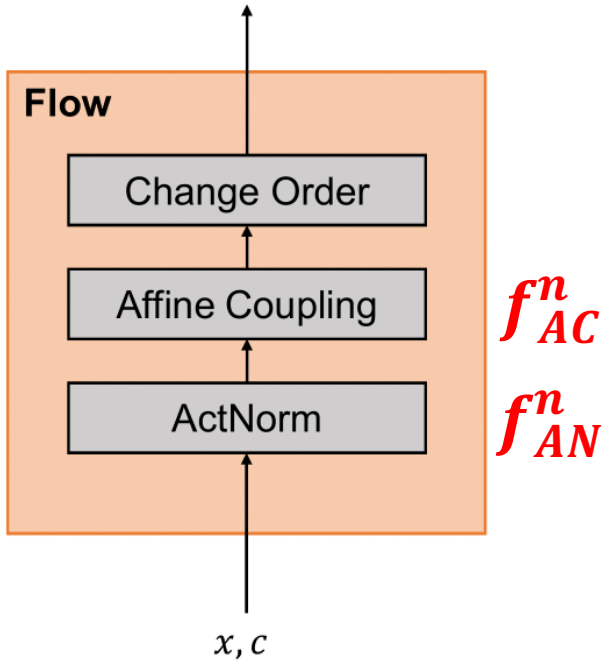
Training phase $$\log p_X(x_{1:T}) = \log p_Z\left(f_\theta(x_{1:T})\right) + \log\det\left(\frac{\partial f_\theta(x)}{\partial x}\right)$$

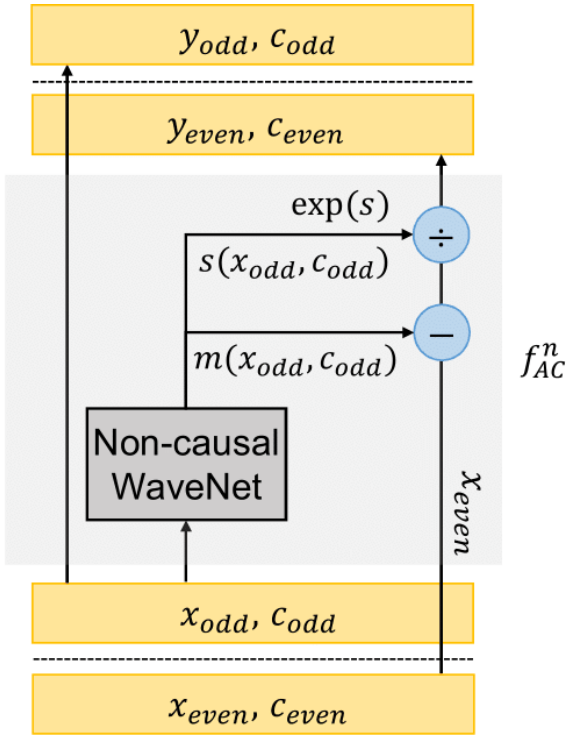
Sampling phase $$z = z_{1:T} \sim P_Z(z) = N(O, I), \qquad \hat{x} = f_\theta^{-1}(z)$$

Both the transformation f_θ and f_θ^{-1} are designed to be computed efficiently
→ Efficient training & Parallel sampling

FloWaveNet







$$\log p_X(x_{1:T}) = \log p_Z\big(f(x_{1:T})\big) + \sum_n \log\det\left(\frac{\partial(f^n_{AC} \cdot f^n_{AN})(x)}{\partial x}\right)$$

Mean Opinion Scores

METHODS	5-SCALE MOS	TEST CLL
GROUND TRUTH	4.67± 0.076	
MOL WAVENET	4.30± 0.110	4.6546
GAUSSIAN WAVENET	4.46± 0.100	4.6526
GAUSSIAN IAF	3.75± 0.159	
FLOWAVENET	3.95± 0.154	4.5457

FloWaveNet ≥ Gaussian IAF

Sampling speed

METHODS	ITER/SEC	SAMPLES/SEC
WAVENET	N/A	172
PARALLEL WAVENET	N/A	500K
GAUSSIAN WAVENET	1.329	44
GAUSSIAN IAF	0.636	470K
FLOWAVENET	0.714	420K

FloWaveNet ≅ Gaussian IAF ≅ Parallel WaveNet >> Autoregressive WaveNet

1000s times faster

Conclusion

- FloWaveNet produces **high quality audio samples** as well as previous distilled models.

- FloWaveNet synthesizes audio samples **in parallel**
  - w/o well pre-trained WaveNet (No distillation!)
  - w/o auxiliary loss terms

Demo page

Code

Poster 6/12 6:30 PM @Pacific Ballroom #2


Thank You!