

# Deep Learning for Audio

## Lecture 6

Pavel Severilov

AI Masters

2024

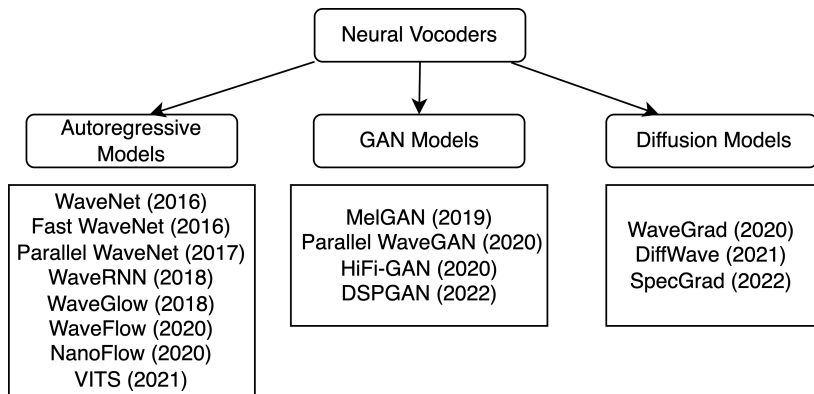
# Outline

1. Neural Vocoder
2. WaveNet
3. Parallel WaveGAN
4. DiffWave
5. VITS

# Outline

1. Neural Vocoders
2. WaveNet
3. Parallel WaveGAN
4. DiffWave
5. VITS

# Neural Vocoders



# Neural Vocoders

Model type	Model	MOS on LJ Speech
<b>Autoregressive</b>	WaveNet	3.68
	WaveRNN	3.96
<b>GAN</b>	MelGAN	3.73
	Parallel WaveGAN	3.99
<b>Diffusion</b>	WaveGrad	3.85
	<b>DiffWave</b>	<b>4.07</b>
	Griffin-Lim	3.68
	<b>Ground Truth</b>	<b>4.10</b>

# Outline

1. Neural Vocoders
2. WaveNet
3. Parallel WaveGAN
4. DiffWave
5. VITS

# WaveNet

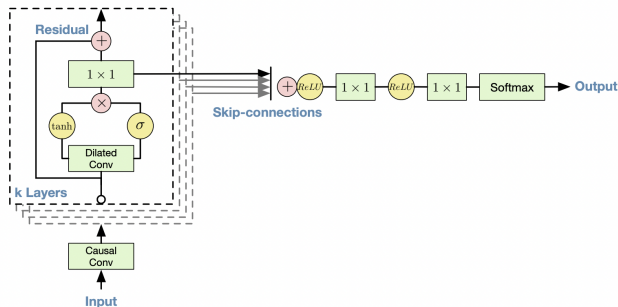
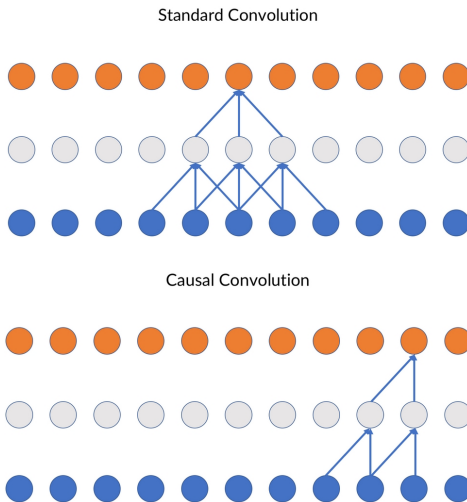


Figure: WaveNet architecture: uses **causal** dilated convolutions

- ▶ The joint probability of a waveform  $x = \{x_1, \dots, x_T\}$ :  
$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t \mid x_{1:t-1})$$
- ▶ Each conditional  $p(x_t \mid x_{1:t-1})$  models the distribution for the timestamp  $t$

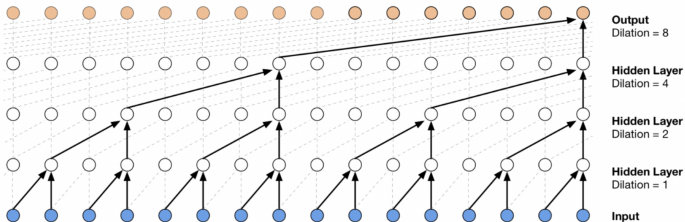
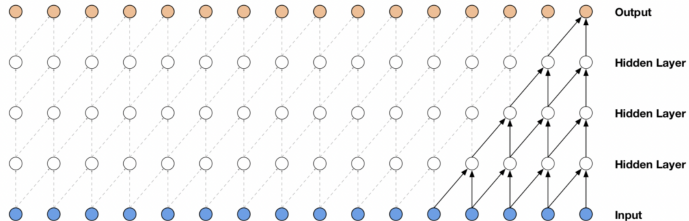
# Causal Convolution



**Figure:** Standard vs causal convolutions. Causal makes convs autoregressive

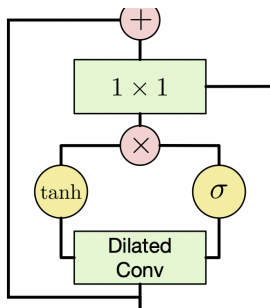


# Dilated Convolution



**Figure:** Non-dilated vs dilated causal convolutions. Dilated convs increase receptive fields

## Conditional gated units

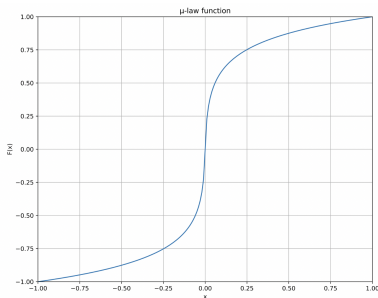


Gated activation unit as used in the gated PixelCNN + condition  $\mathbf{y}$

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}).$$

$*$  – convolution,  $\odot$  – element-wise multiplication,  $\sigma(\cdot)$  – sigmoid function,  $f$  and  $g$  – filter and gate, respectively,  $W$  – learnable convolution filter,  $V$  – learnable linear projection

# Mu Law Encoding



- ▶ Raw audio  $\sim$  16-bit integer values  $\Rightarrow$  softmax layer need to output 65,536 probabilities per timestep
- ▶ Solution: apply a  $\mu$ -law transformation to the data, and then quantize it to 256 possible values:

$$f(x_t) = \text{sign}(x_t) \frac{\ln(1 + \mu |x_t|)}{\ln(1 + \mu)}, \quad -1 < x_t < 1, \mu = 255$$

# Outline

1. Neural Vocoders
2. WaveNet
3. Parallel WaveGAN
4. DiffWave
5. VITS

# Parallel WaveGAN

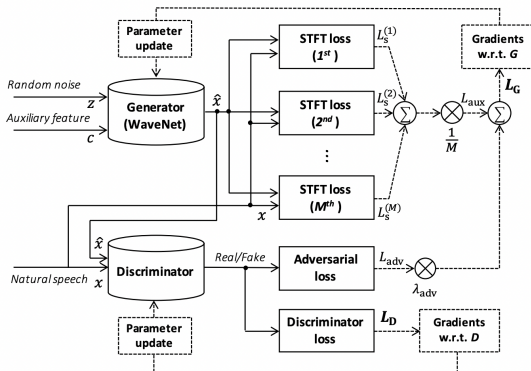
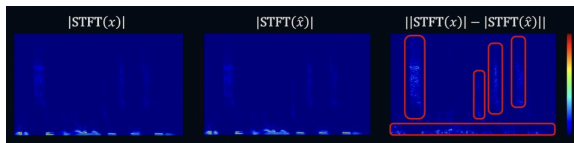


Figure: Parallel WaveGAN

Yamamoto et al., *Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram*, IEEE ICASSP, 2020

# STFT Loss



$$\mathcal{L}_s(G) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}), \mathbf{x} \sim p_{data}} [\mathcal{L}_{sc}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{mag}(\mathbf{x}, \hat{\mathbf{x}})]$$

$$\mathcal{L}_{sc}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{|||STFT(\mathbf{x})| - |STFT^T(\hat{\mathbf{x}})|||}{|||STFT(\mathbf{x})|||_F}$$

$$\mathcal{L}_{mag}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \|\log |STFT(\mathbf{x})| - \log |STFT(\hat{\mathbf{x}})|\|_1$$

$$\mathcal{L}_{aux}(G) = \frac{1}{M} \sum_{m=1}^M \mathcal{L}_s^{(m)}(G), M - \text{number of STFT losses}$$

Takaki et al., *STFT Spectral Loss for Training a Neural Speech Waveform Model*,  
IEEE ICASSP, 2019

# GAN Loss

- Discriminator Loss:

$$\mathcal{L}_D(G, D) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [(1 - D(\mathbf{x}))^2] + \mathbb{E}_{\mathbf{z} \sim N(0, I)} [D(G(\mathbf{z}))^2]$$

- Generator Loss

$$L_{\text{adv}}(G, D) = \mathbb{E}_{\mathbf{z} \sim N(0, I)} [(1 - D(G(\mathbf{z})))^2]$$

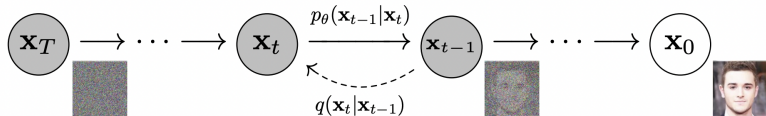
$$L_G(G, D) = L_{\text{aux}}(G) + \lambda_{\text{adv}} L_{\text{adv}}(G, D)$$

# Outline

1. Neural Vocoders
2. WaveNet
3. Parallel WaveGAN
4. DiffWave
5. VITS



# Diffusion models idea



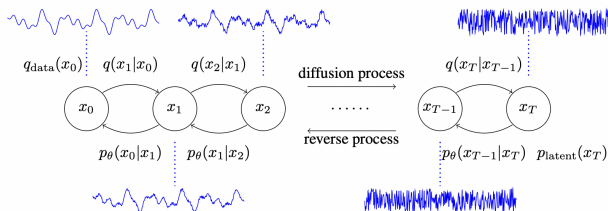
- **Diffusion probabilistic model:** parameterized Markov chain from data  $x_0$  to the latent variable  $x_T$

$$q(x_1, \dots, x_T | x_0) = \prod_{t=1}^T q(x_t | x_{t-1})$$

- **Reverse process:** Markov chain from  $x_T$  to  $x_0$  parameterized by  $\theta$ :

$$p_{\text{latent}}(x_T) = \mathcal{N}(0, I), \quad p_{\theta}(x_0, \dots, x_{T-1} | x_T) = \prod_{t=1}^T p_{\theta}(x_{t-1} | x_t)$$

# DiffWave



- **Sampling:** reverse process  $x_T \sim \mathcal{N}(0, I)$ ,  $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$  for  $t = T, T-1, \dots, 1$ .  $x_0$  – sampled data.
- **Training:**  $p_\theta(x_0) = \int p_\theta(x_0, \dots, x_{T-1}|x_T) \cdot p_{\text{latent}}(x_T) \mathbf{d}x_{1:T}$  (likelihood) is intractable to calculate  $\Rightarrow$  model trained by maximizing its variational lower bound (ELBO):

$$\begin{aligned} \mathbb{E}_{q_{\text{data}}(x_0)} \log p_\theta(x_0) &\geq \\ &\geq \mathbb{E}_{q(x_0, \dots, x_T)} \log \frac{p_\theta(x_0, \dots, x_{T-1}|x_T) \cdot p_{\text{latent}}(x_T)}{q(x_1, \dots, x_T|x_0)} := \text{ELBO} \end{aligned}$$

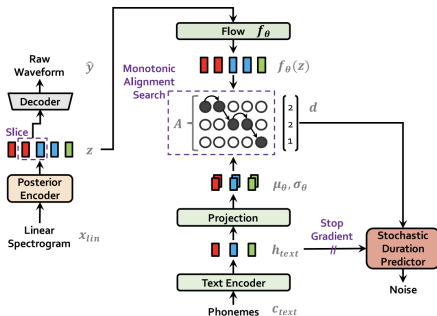
## DiffWave: advantages

- ▶ **Non-autoregressive**: much faster than WaveNet
- ▶ **Compact model**: smaller footprint than flow-based models
- ▶ No auxiliary losses in training (e.g., spectrogram-based losses): **no mode collapse** like in GANs/VAEs

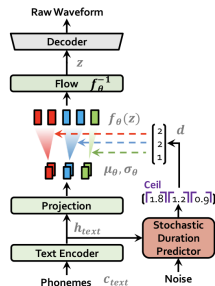
# Outline

1. Neural Vocoders
2. WaveNet
3. Parallel WaveGAN
4. DiffWave
5. VITS

# VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech)



(a) Training procedure



(b) Inference procedure

- ▶ conditional VAE with the objective of maximizing ELBO (conditional prior: green blocks)
- ▶ Inference: text  $\Rightarrow$  duration prediction  $\Rightarrow$  cascade of the flow module and HiFi-GAN decoder  $\Rightarrow$  waveform

Kim, Jaehyeon et al., *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*, ICML, 2021

# VITS blocks

- ▶ Posterior encoder: non-causal WaveNet residual blocks
- ▶ Prior encoder: text encoder that processes the input phonemes  $c_{text}$  and a normalizing flow  $f_\theta$  that improves the flexibility of the prior distribution
- ▶ text encoder: transformer encoder with relative positional representation
- ▶ Stochastic duration predictor: flow-based generative model
- ▶ Normalizing flow: stack of affine coupling layers consisting of a stack of WaveNet residual blocks
- ▶ Decoder: HiFi-GAN V1 generator