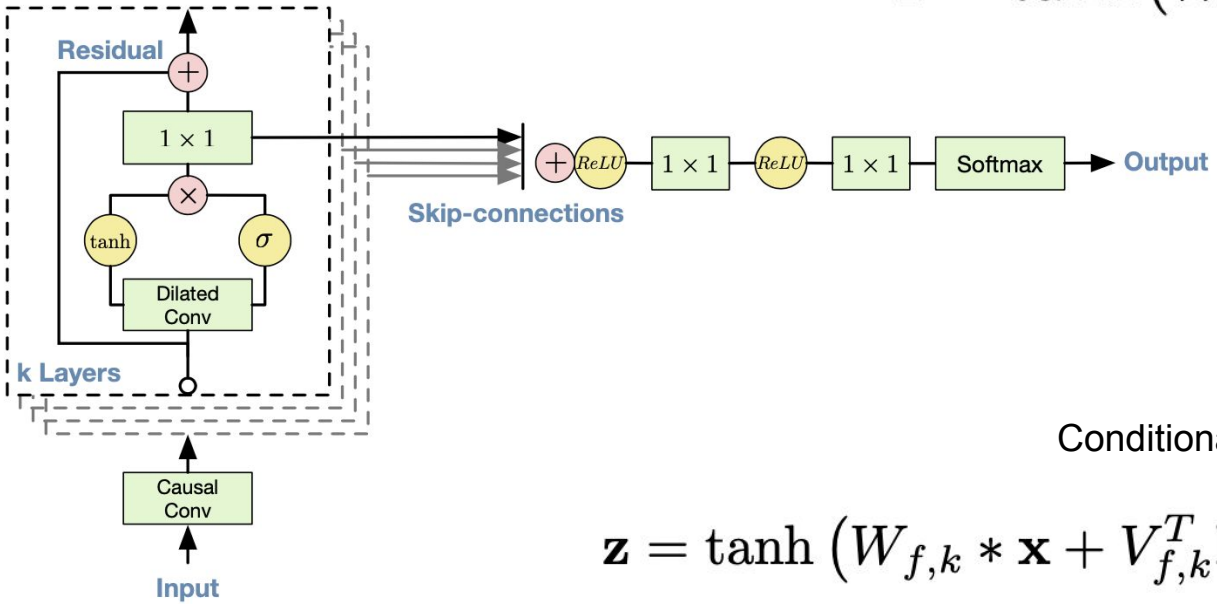
TTS: Vocoders

History lesson aka WaveNet

autoregressive model i.e. $p(x) = \prod_{t=1}^{T} p(x_t|x_1 \ldots x_{t-1})$

Gated Activation Unit

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$



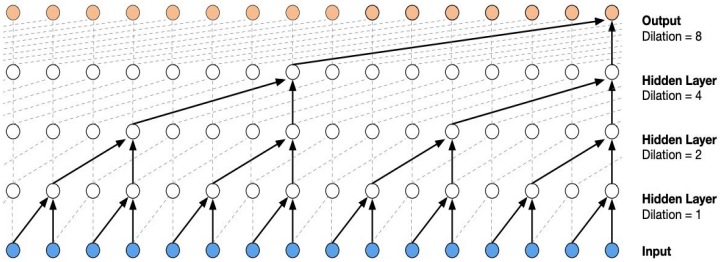
Conditional Gated Activation Unit

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^{T}\mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^{T}\mathbf{h}\right)$$

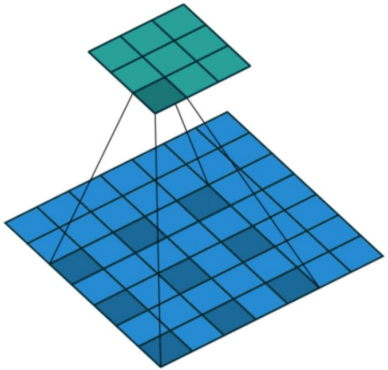
Dilated Convolution:

increase receptive field to better process long-term dependencies

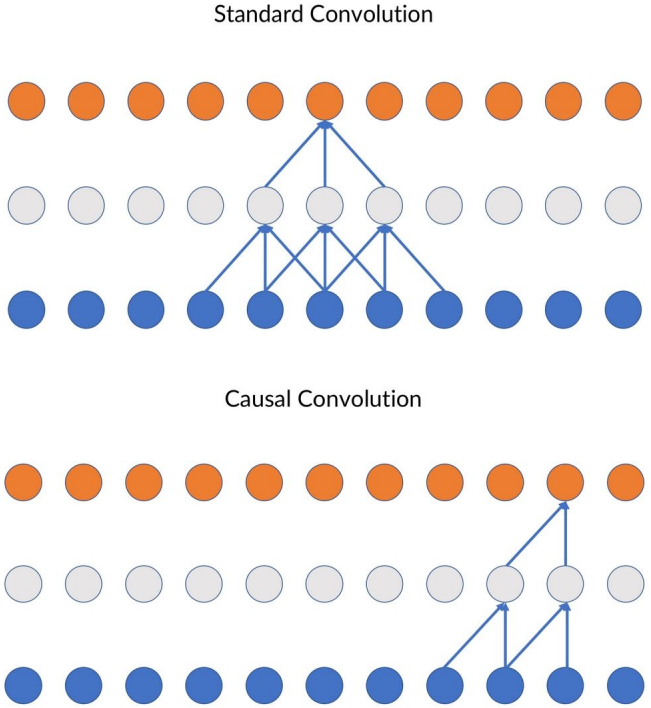
1D



2D



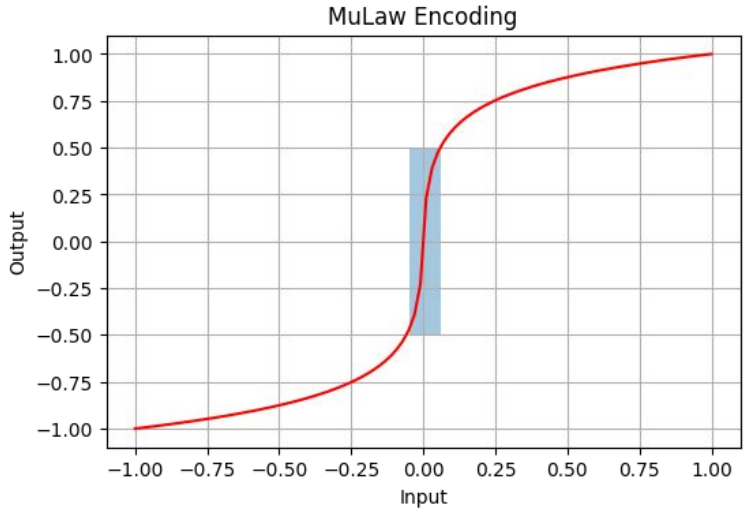
Causal Convolution:



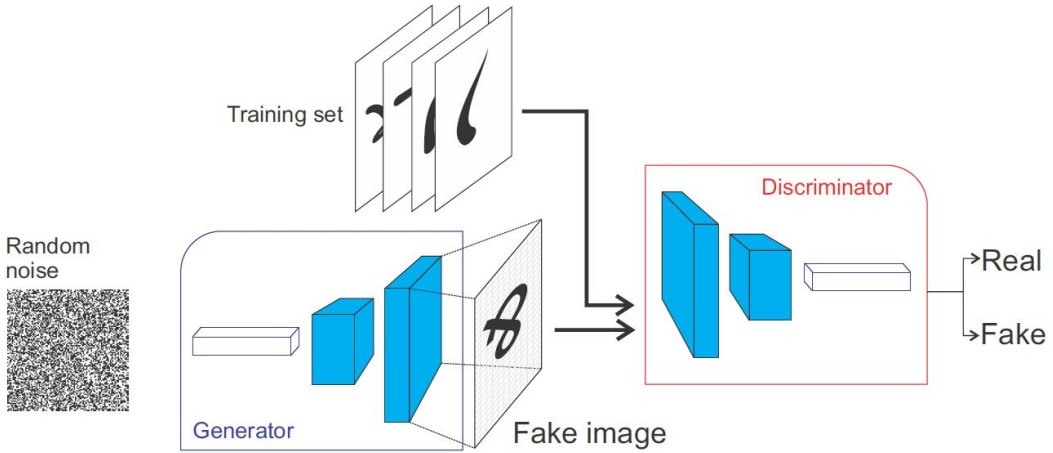
Mu Law Encoding

$$f(x_t) = \text{sign}(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

$$\mu = 255 \qquad -1 \leq x_t \leq 1$$



Generative Adversarial Networks (GANs)



1) Vanilla adversarial loss:

G - $log(D(G(z_i)))$

D - $logD(x_i) + log(1 - D(G(z_i)))$

2) Least Squares loss: (LS-GAN)

G - $(D(G(z_i)) - 1)^2$

D - $(D(x_i) - 1)^2 + (D(G(z_i)))^2$

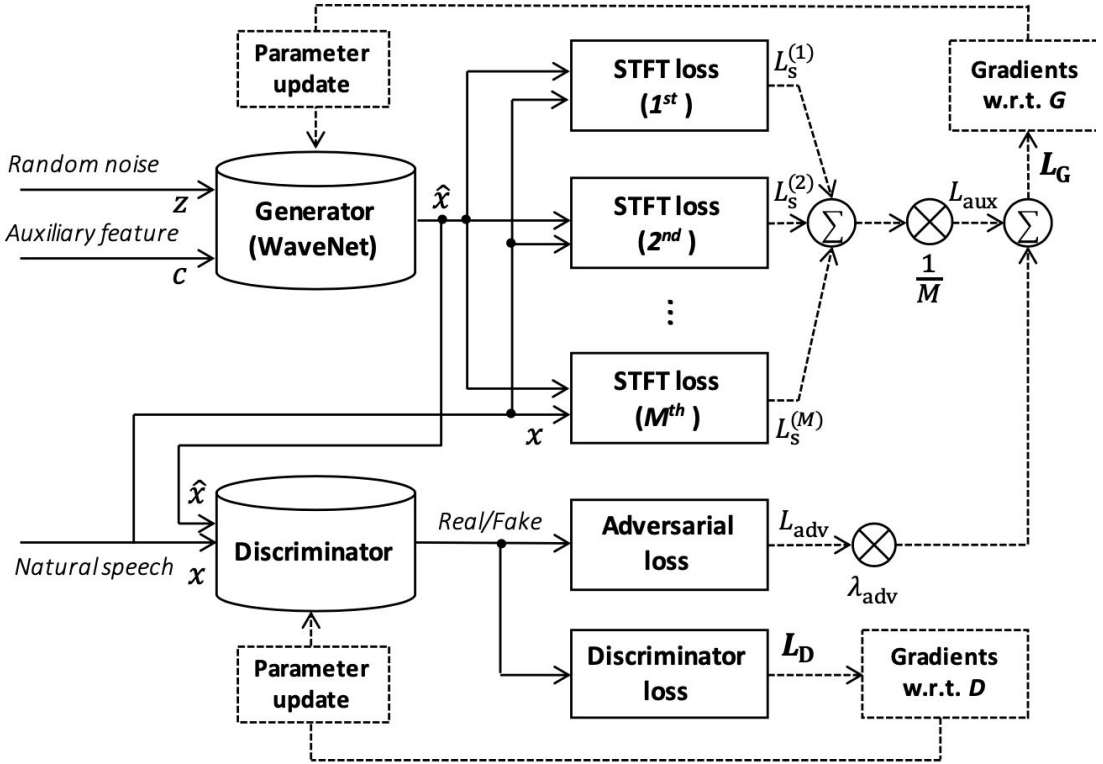
3) Feature Matching loss:

G - $\sum_{i=1}^{T} \frac{1}{N_i} ||D_k^{(i)}(x_j) - D_k^{(i)}(G(z_j))||_1$

4) Markovian loss (aka PatchGan)

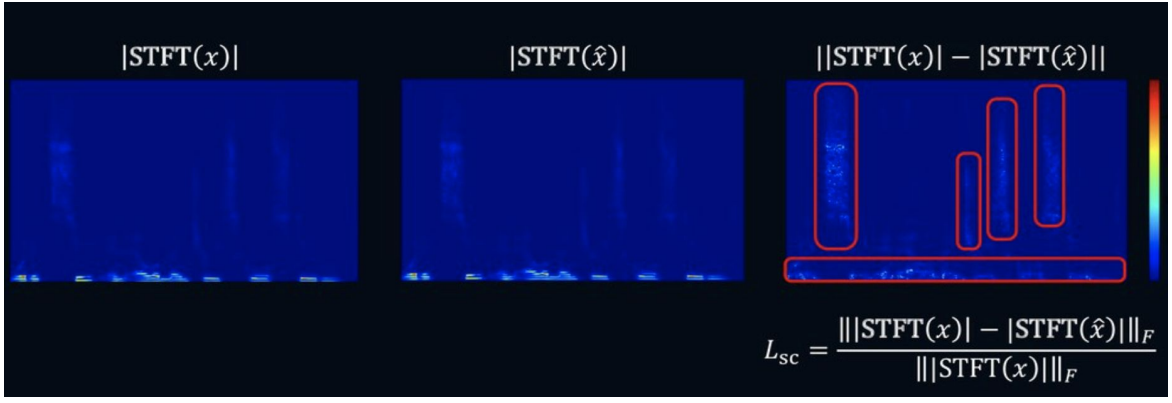
Parallel WaveGAN

- 2 stage training
- WaveNet generator
- multi resolution stft loss

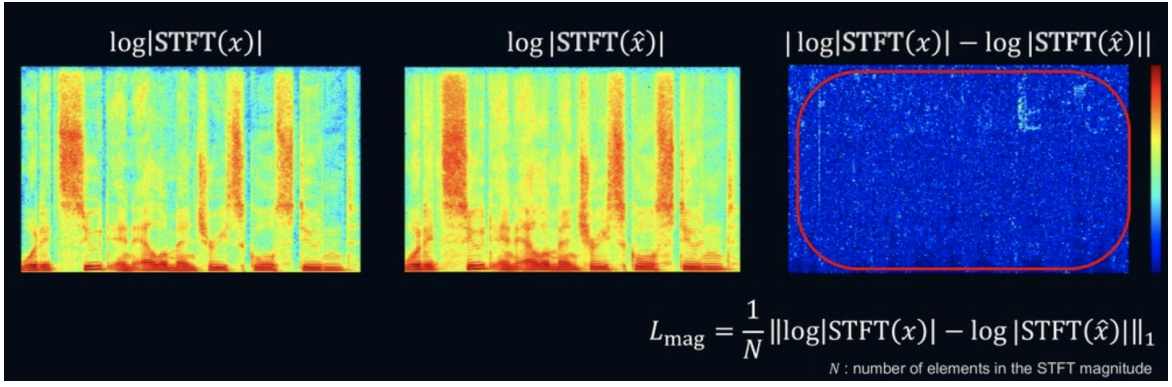


STFT Loss

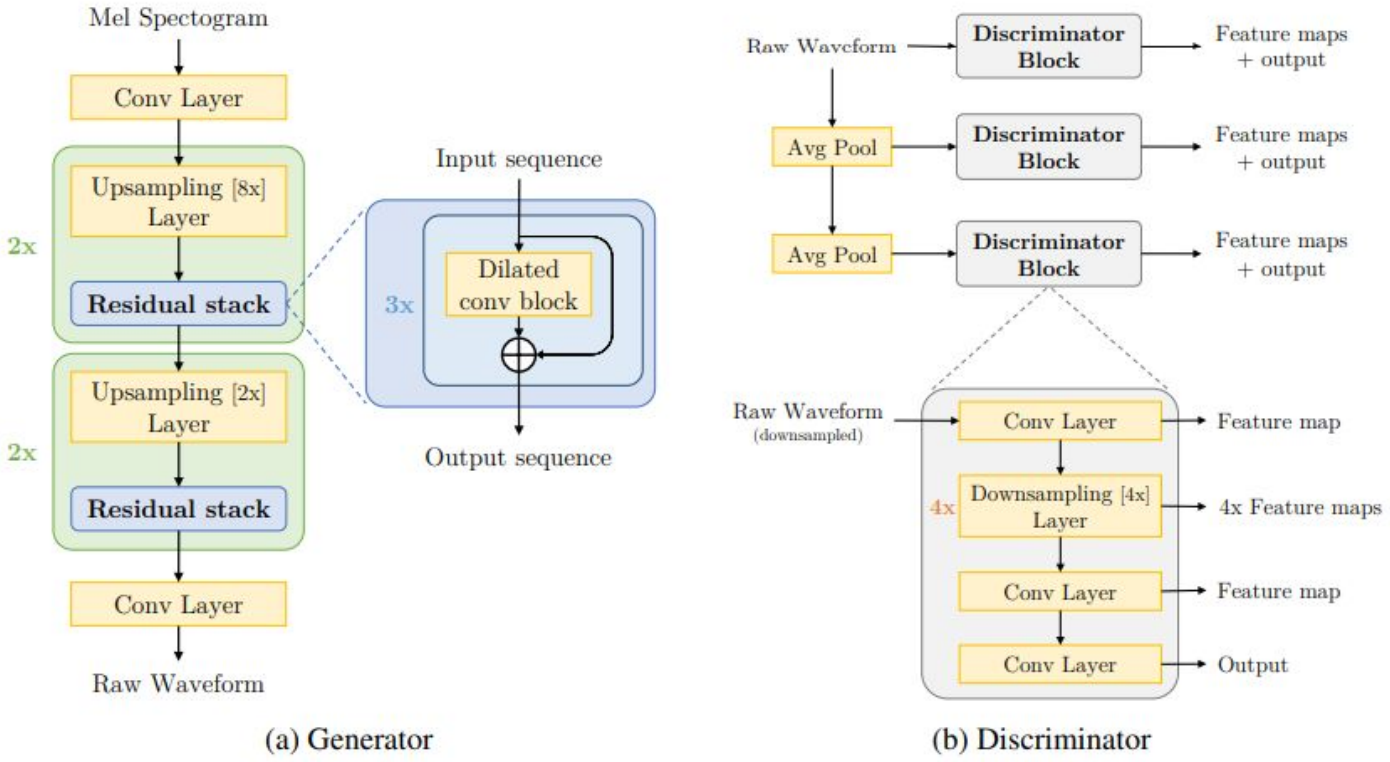
Spectral Convergence part



Log scale STFT magnitude part

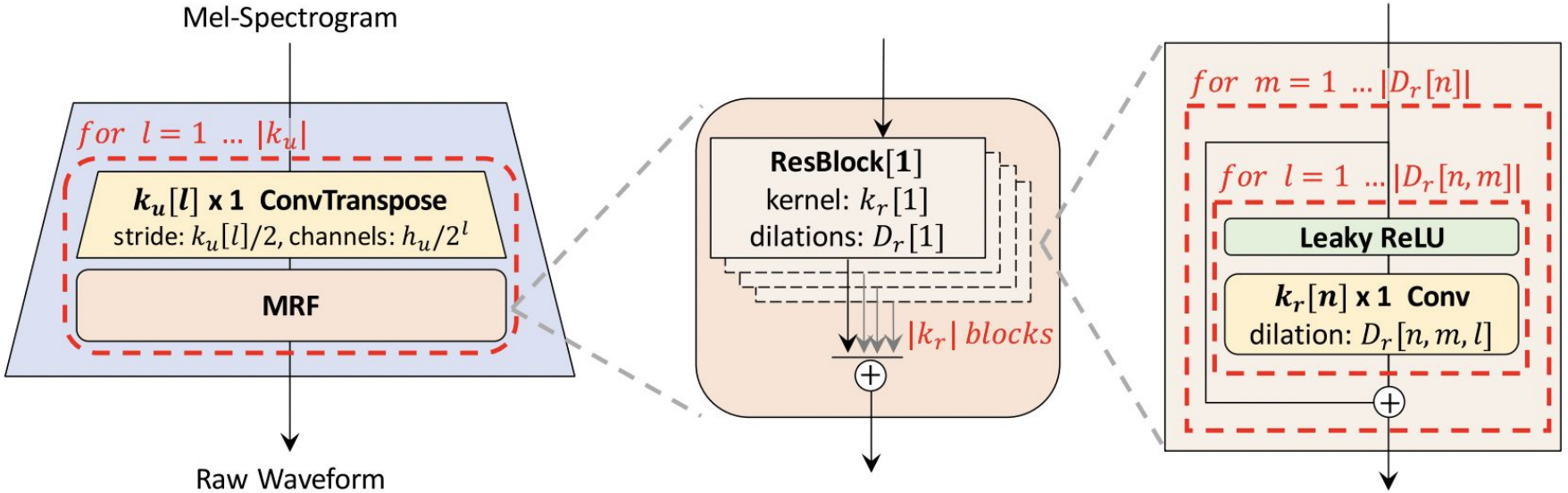


MelGan


Mel Spectogram
Conv Layer
Upsampling [8x] Layer
2x
Residual stack
Upsampling [2x] Layer
2x
Residual stack
Conv Layer
Raw Waveform
Input sequence
3x
Dilated conv block
Output sequence
(a) Generator
Raw Waveform
Discriminator Block
Feature maps + output
Avg Pool
Discriminator Block
Feature maps + output
Avg Pool
Discriminator Block
Feature maps + output
Raw Waveform (downsampled)
Conv Layer
Feature map
4x
Downsampling [4x] Layer
4x Feature maps
Conv Layer
Feature map
Conv Layer
Output
(b) Discriminator

Hifi-GAN

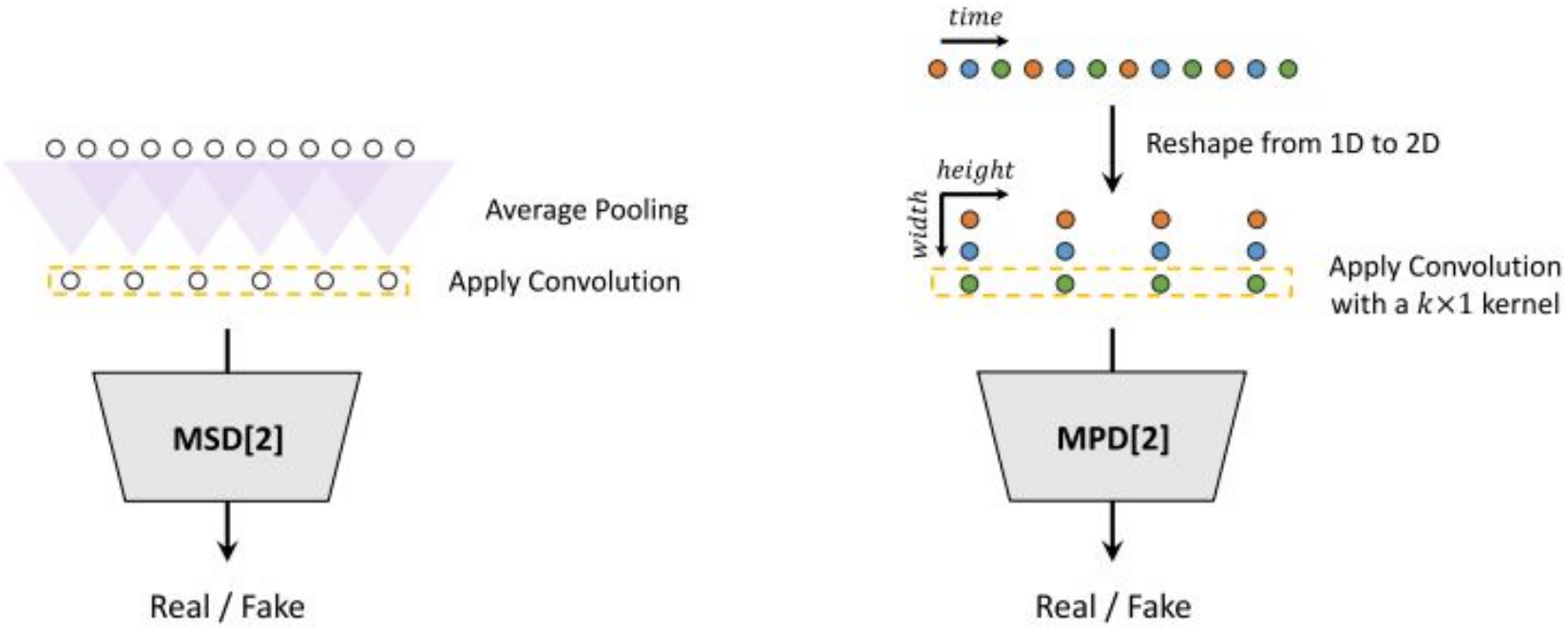
Generator:



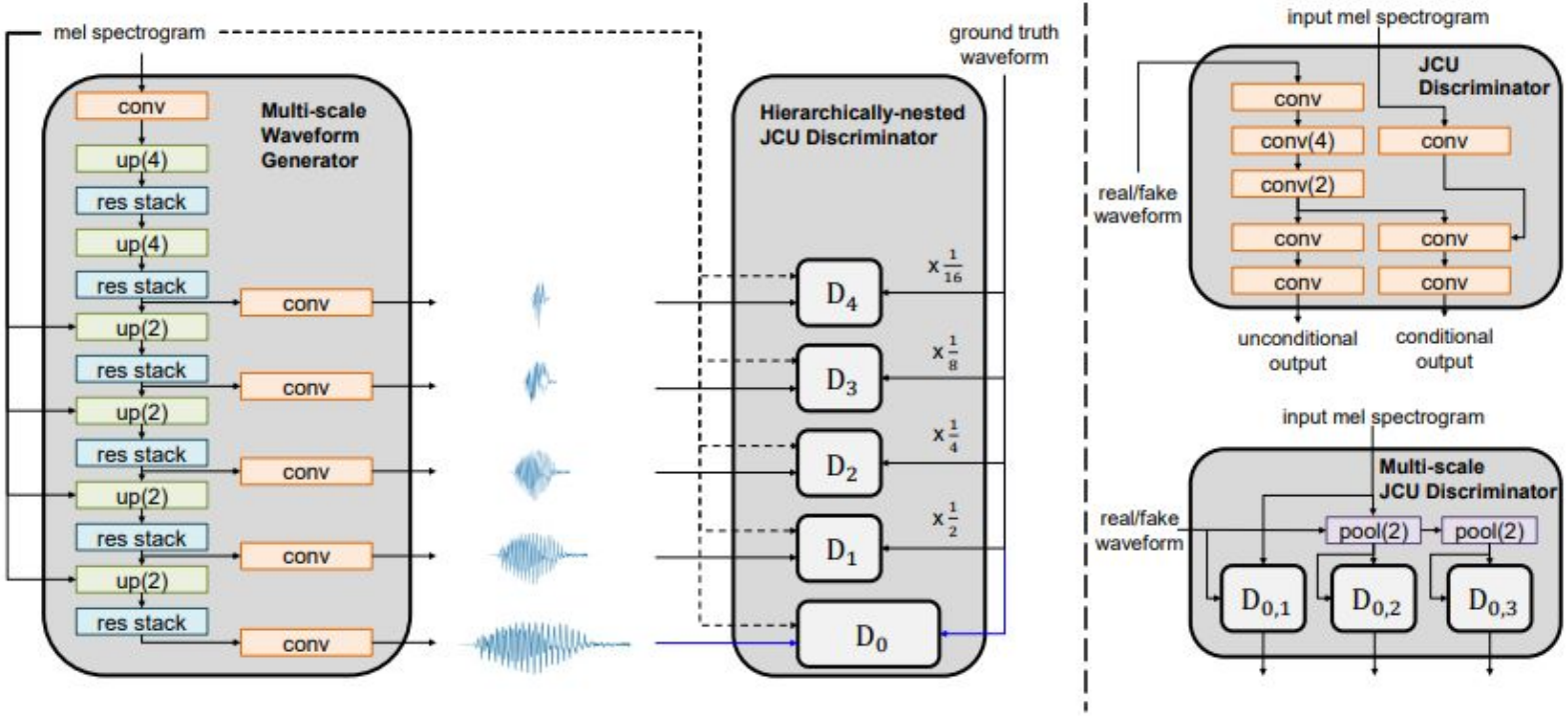
h_u	k_u	k_r	D_r
512	[16, 16, 4, 4]	[3, 7, 11]	[[1, 1], [3, 1], [5, 1]] ×3

Hifi-GAN

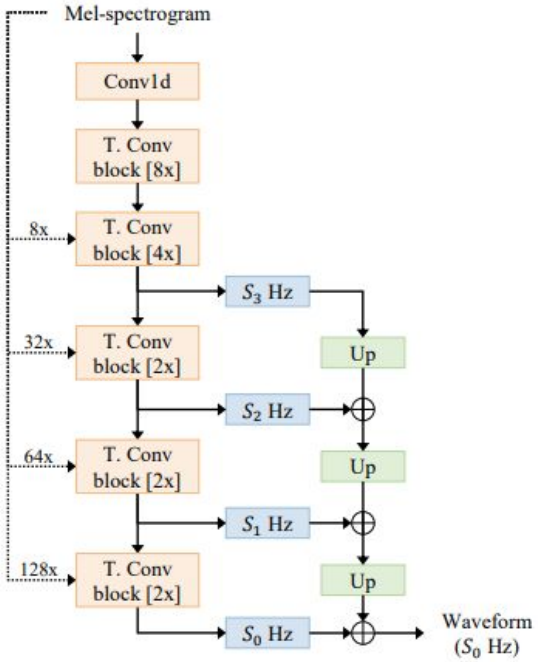
Discriminator:



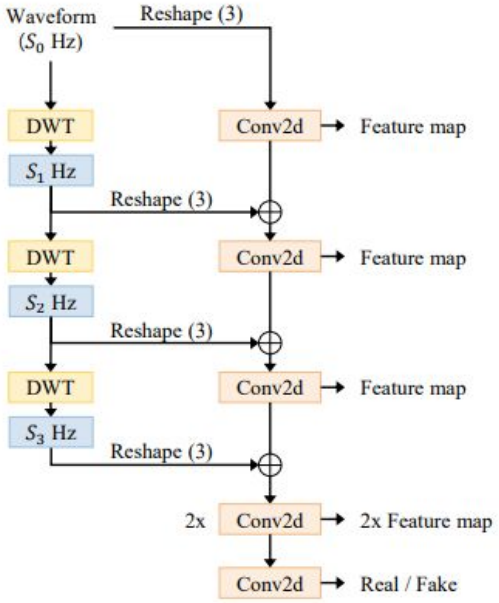
VocGAN


mel spectrogram
conv
up(4)
res stack
up(4)
res stack
up(2)
res stack
up(2)
res stack
up(2)
res stack
up(2)
res stack
Multi-scale Waveform Generator
conv
conv
conv
conv
conv
ground truth waveform
Hierarchically-nested JCU Discriminator
D_4
D_3
D_2
D_1
D_0
$x\frac{1}{16}$
$x\frac{1}{8}$
$x\frac{1}{4}$
$x\frac{1}{2}$
input mel spectrogram
real/fake waveform
JCU Discriminator
conv
conv(4)
conv(2)
conv
conv
conv
conv
conv
unconditional output
conditional output
input mel spectrogram
real/fake waveform
Multi-scale JCU Discriminator
pool(2)
pool(2)
$D_{0,1}$
$D_{0,2}$
$D_{0,3}$

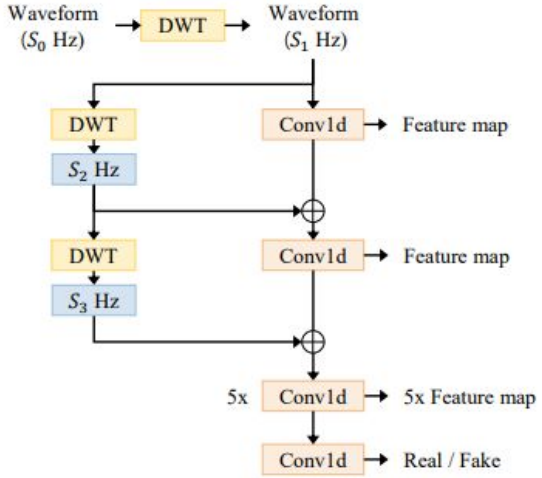
Fre-GAN



(a) RCG

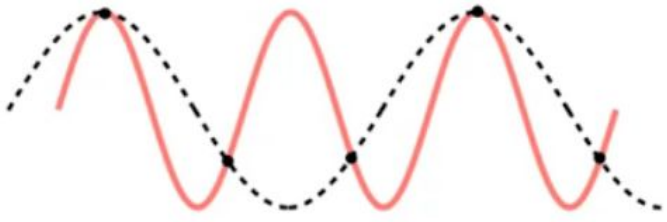


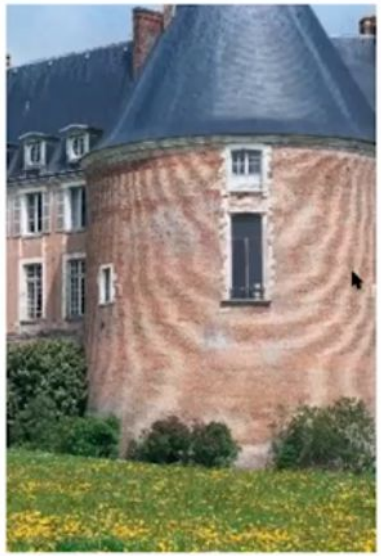
(b) RPD

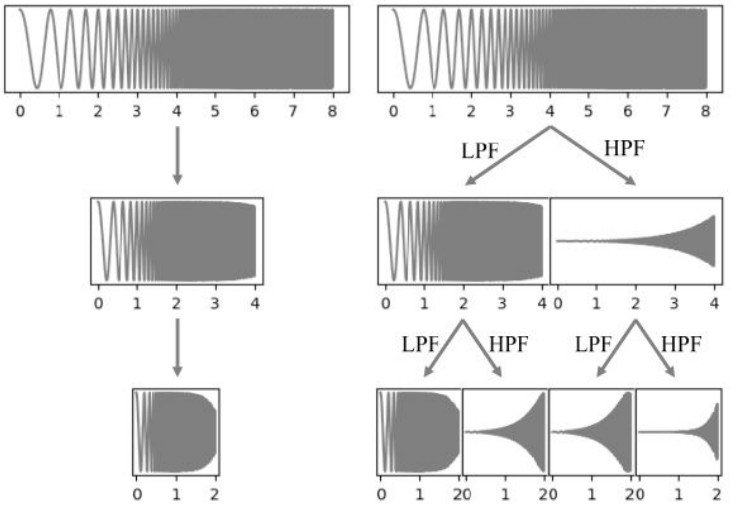


(c) RSD

Discrete Wavelet Transform (DWT)







(a) Average Pooling (b) Discrete Wavelet Transform

Figure 1: *Comparison of Average Pooling (AP) and Discrete Wavelet Transform (DWT). Here, LPF and HPF refer to Low-Pass and High-Pass Filter, respectively. In this example, an up-chirp signal whose frequency increases from 0 Hz at time $t = 0$ to 150 Hz at time $t = 8$ is downsampled by AP and DWT.*

Other vocoders:

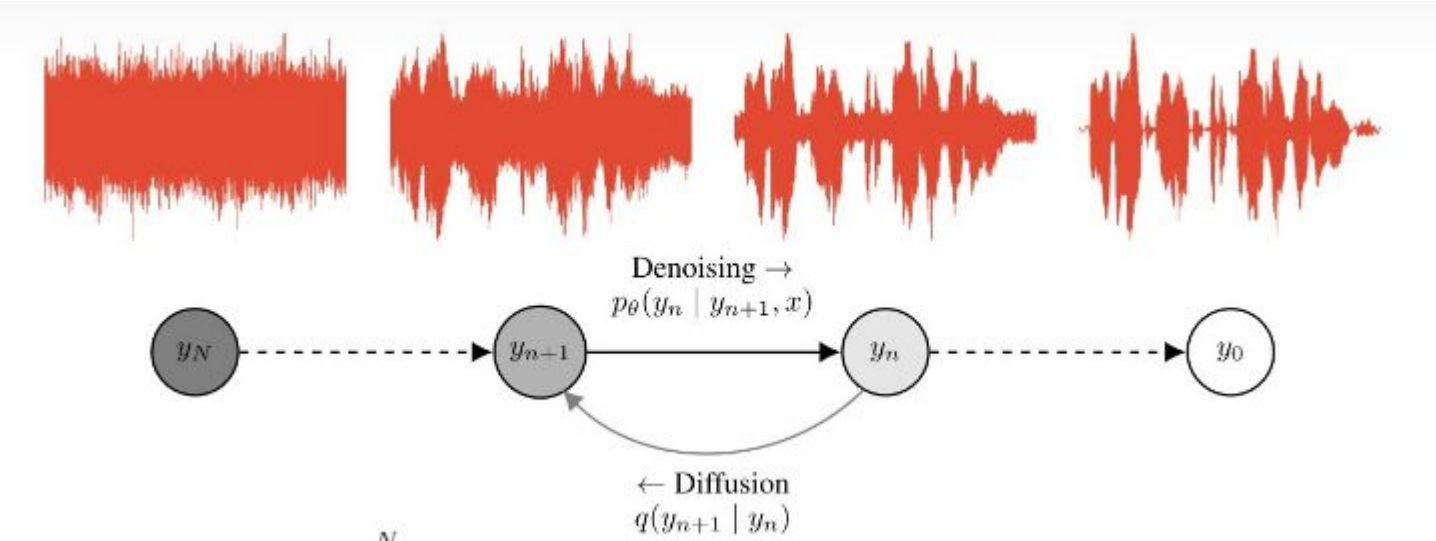
FLOWs:

- Parallel WaveNet
- WaveGlow
- WaveFlow
- NanoFlow

Denoising Diffusion Probabilistic Models (DDPM):

- Diff-TTS
- more on DDPM

Denoising Diffusion Probabilistic Models



$$q(y_1, ..., y_N|y_0) := \prod_{n=1}^{N} q(y_n|y_{n-1})$$

$$q(y_n|y_{n-1}) := \mathcal{N}(y_t; \sqrt{1-\beta_t}x_{n-1}, \beta_t I)$$

more on DDPMs