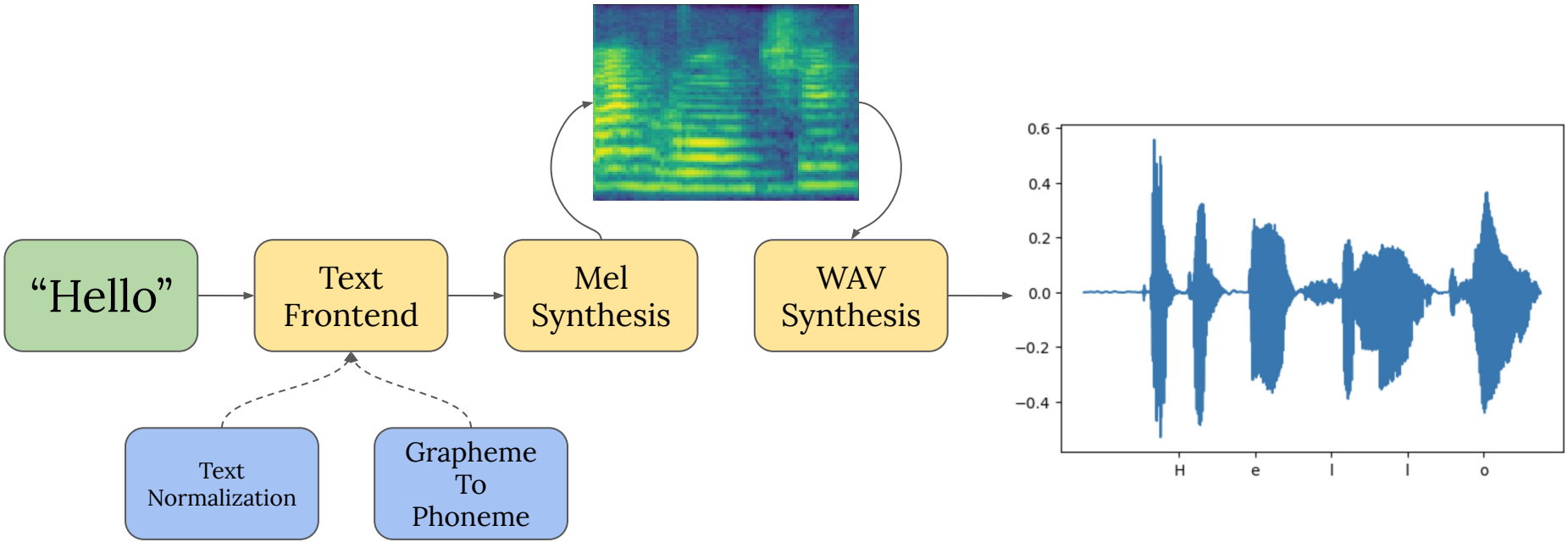
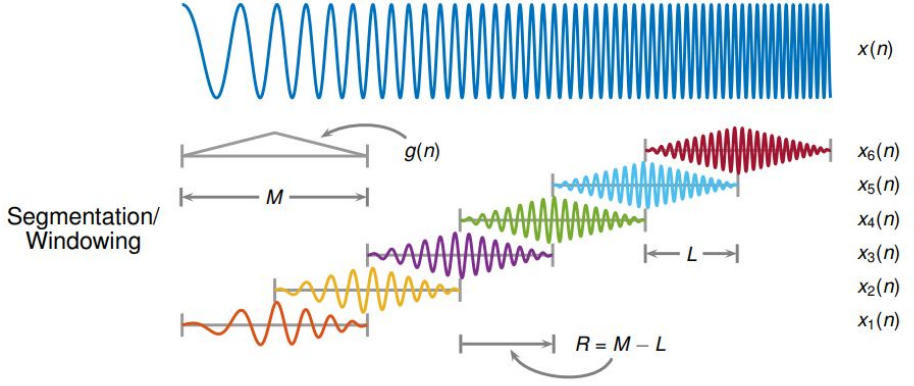
Generative Models for Speech Synthesis

Markovich Alexander

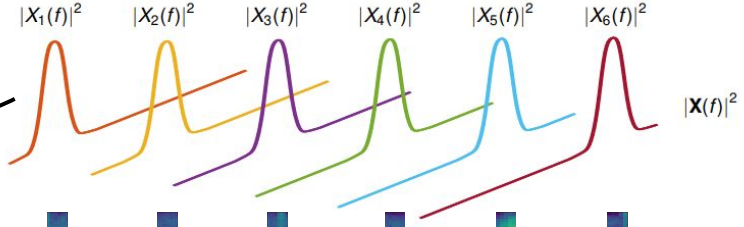
What is a TTS in general?



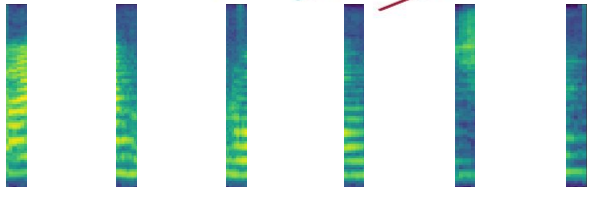
What is MelSpectrogram?

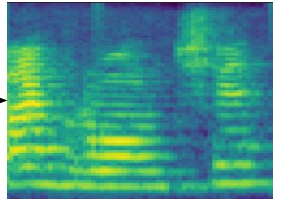


DFT



log(MelScale(x))

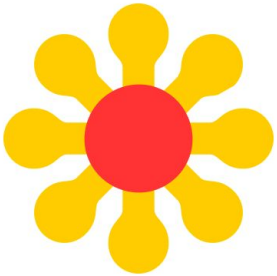


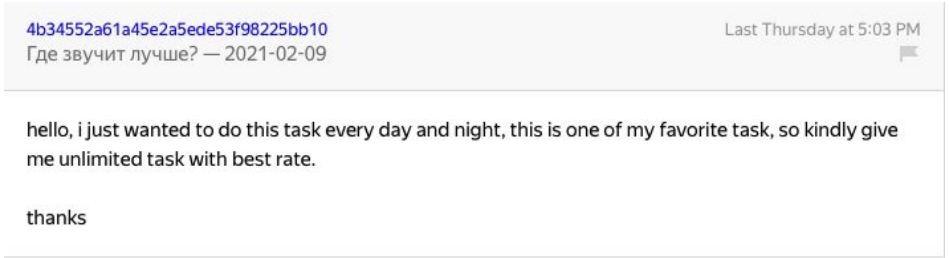


How to Measure Quality?

- There’s no correct answer
- Subjective perception
- A lot of types of mistakes
- The only solution is MOS

amazon mechanical turk





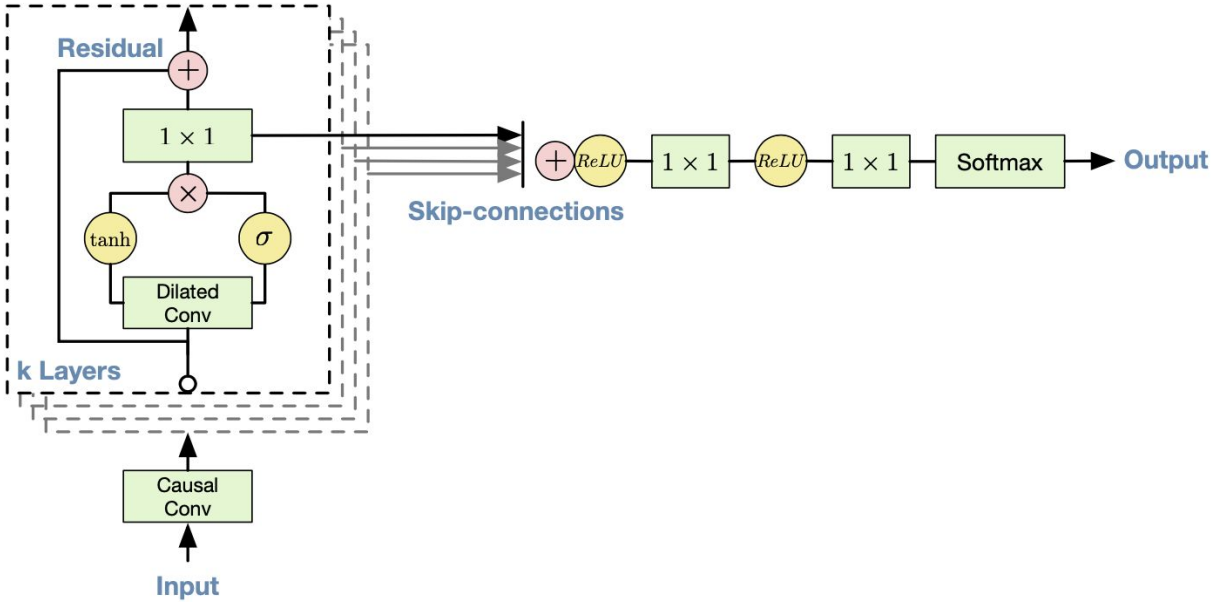
Neural Vocoders

- WaveNet
- Clari & Parallel WaveNet
- WaveGlow
- WaveFlow
- NanoFlow
- MelGAN
- Parallel WaveGAN

WaveNet

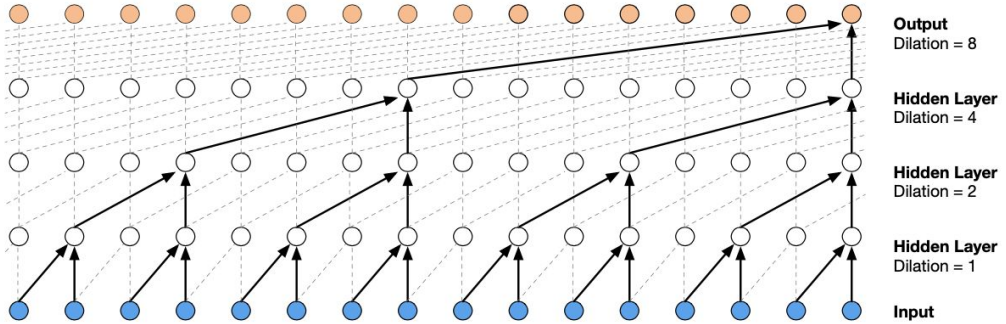
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

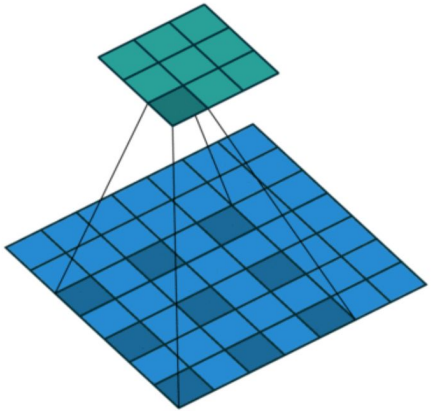
$$p(x_t | x_i, \ldots, x_{t-1}) \sim \mathrm{Cat}(\pi_\theta)$$



Dilated Convolution

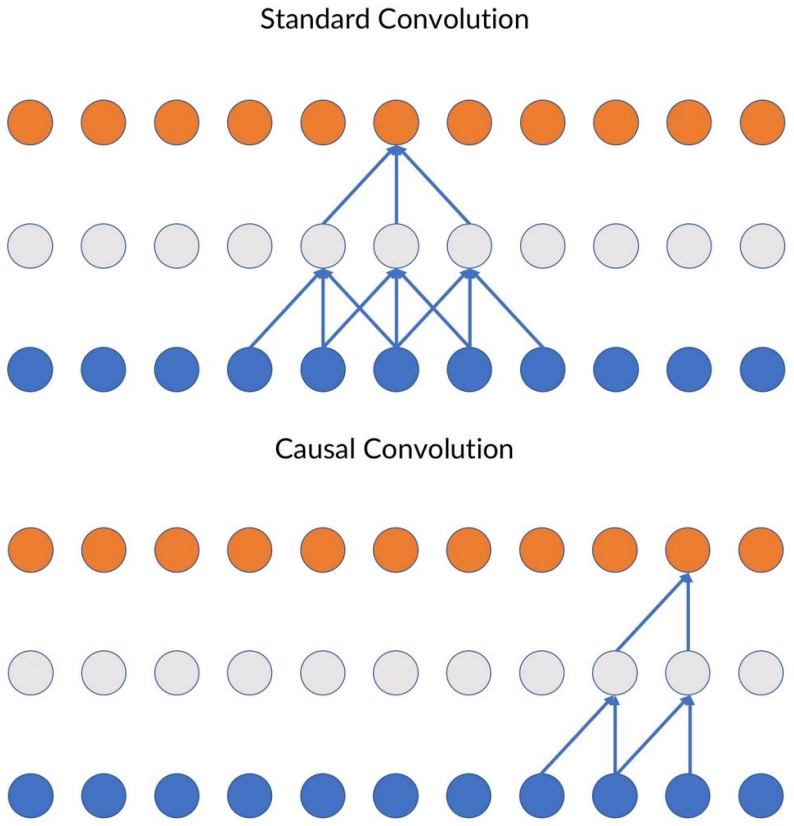
- Increase receptive field
- Allow modeling long time dependencies





Causal Convolution

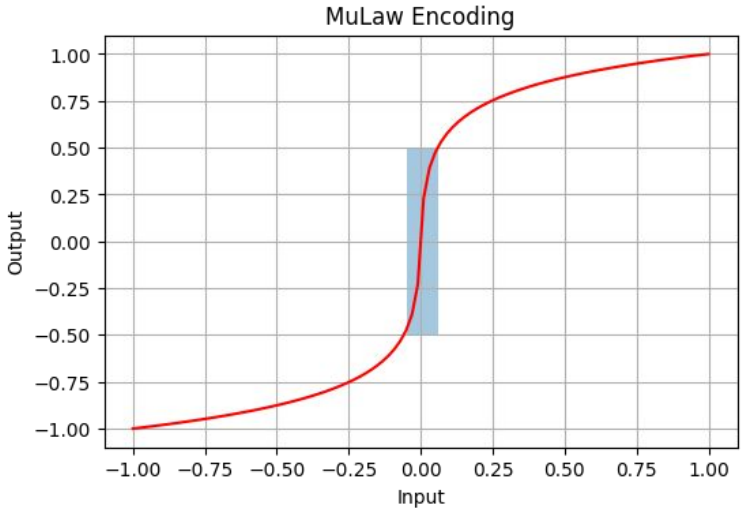
- $p(x_{t+1} \mid x_1, \ldots, x_t)$
- Don’t use padding in Conv
- Use a separate Pad



Mu Law Encoding

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

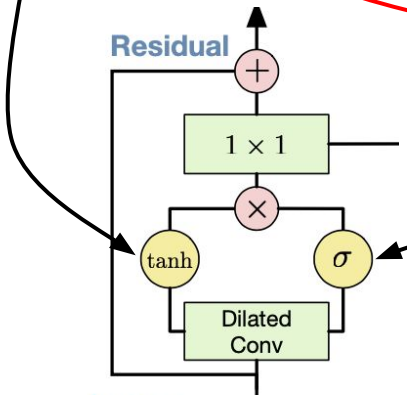
- 16-bit WAV contain 2^16 values
- Softmax will die :(
- Human hearing on a **logarithmic** scale
- **Low-amplitude** sounds in **high** resolution
- **High-amplitude** sounds in **low** resolution

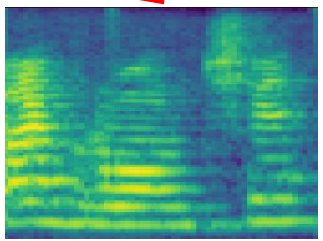


http://www.young-engineering.com/docs/YoungEngineering_ALaw_and_MuLaw_Companding.pdf

(Condition) Gated Mechanism

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





But how do we align the WAV and Mel?

Upsample is our everything!

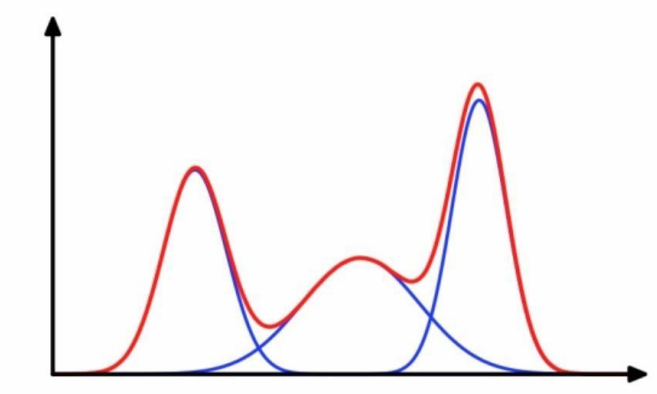

```
1  torch.nn.Upsample(hop_size, mode='linear')
```

```
Upsample(size=256, mode=linear)
```

```
1  torch.nn.ConvTranspose1d(..., stride=hop_size)
```

What about a loss function?

- Categorical distribution
- Normal distribution
- Logistic distribution
- Mixture of Normals or Logistics
- Use torch.distributions :)



Sampling is dangerous!

Clari & Parallel WaveNet

- Gaussian IAF based on WaveNet
- $\boldsymbol{z} \sim \text{Logistic}(0, I)$ or $\mathcal{N}(0, I)$
- Shift (**mu**) and scale (**sigma**) are modeled by WaveNet
- Probability Density Distillation
- STFT Loss

$$z_t = \frac{x_t - \mu(z_{<t}; \boldsymbol{\vartheta})}{\sigma(z_{<t}, \vartheta)}$$

$$x_t = z_t \cdot \sigma(z_{<t}; \boldsymbol{\vartheta}) + \mu(z_{<t}; \boldsymbol{\vartheta})$$

Probability Density Distillation

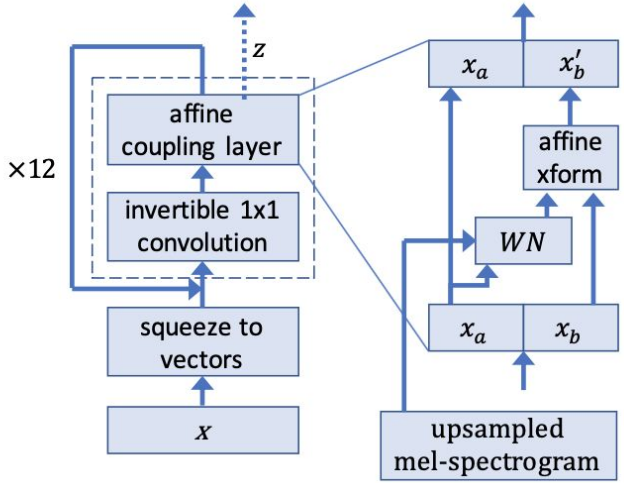
- Sample $\boldsymbol{z} \sim \text{Logistic}(0, I)$ or $\mathcal{N}(0, I)$
- Pass $\mathbf{z}$ into **IAF** and obtain $q(x_t \mid z_{<t}; \boldsymbol{\vartheta}) \sim \mathcal{N}$ or Logistic
- Calculate KL Divergence between **Student** and **Autoregressive Teacher:**

$$D_{\text{KL}}(P_S \| P_T) = H(P_S, P_T) - H(P_S)$$

- There is often a **closed-form** formula

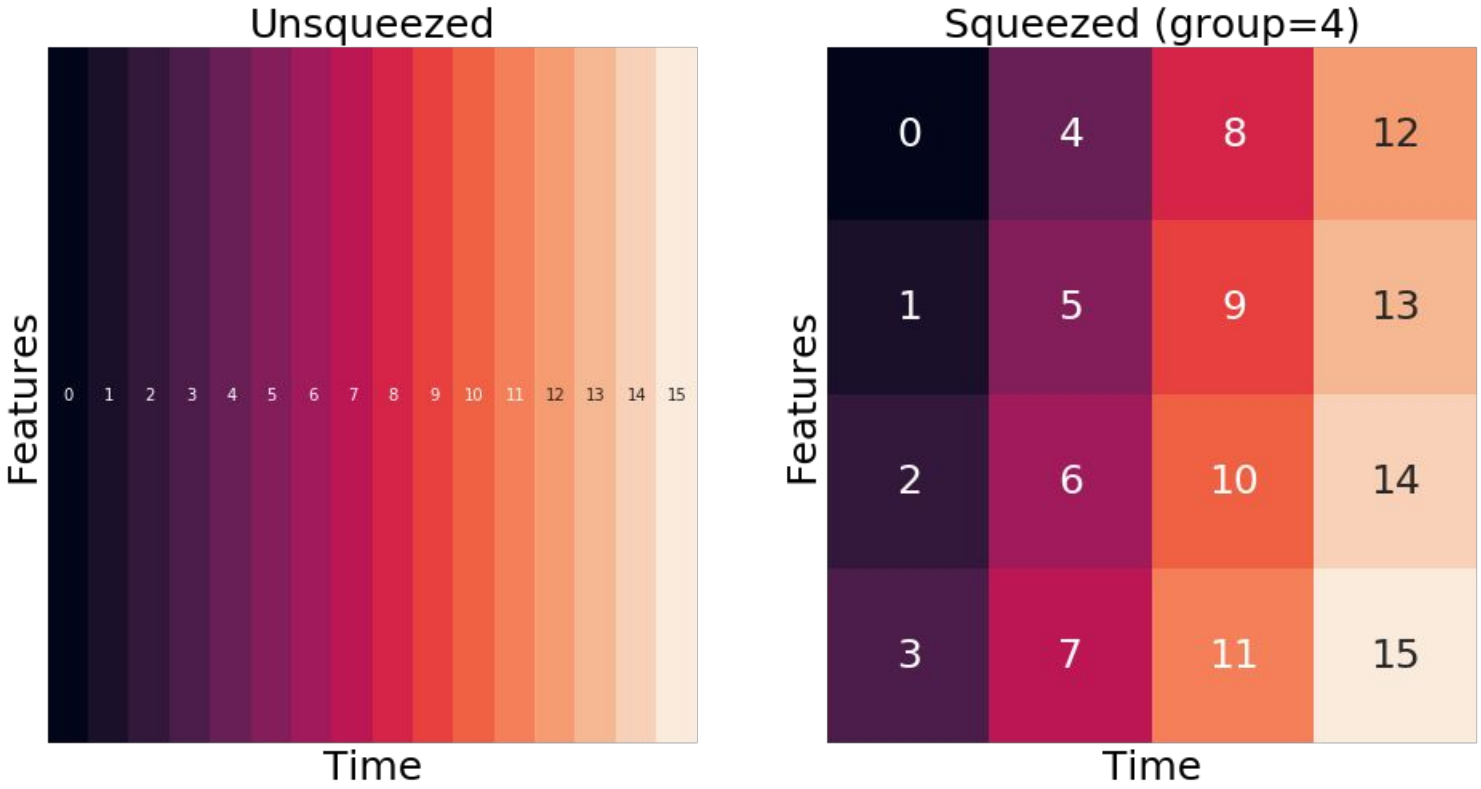
WaveGlow

- Combine insights from **Glow** and **WaveNet**
- Squeeze Operation
- Affine Couling Layer
- 1x1 Invertible Convolution
- Early outputs



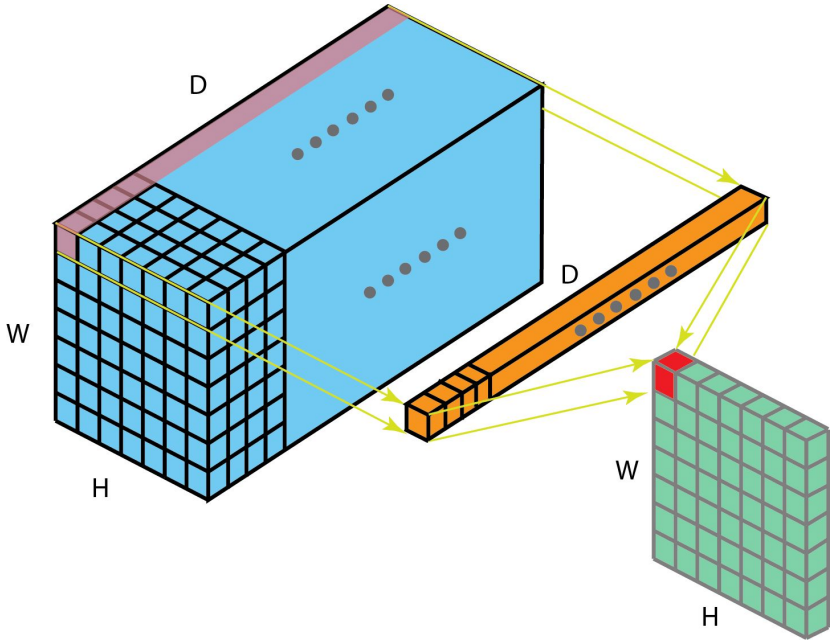
$$\boldsymbol{x}_a, \boldsymbol{x}_b = \text{split}(\boldsymbol{x})$$
$$(\log \boldsymbol{s}, \boldsymbol{t}) = \text{WaveNet}(\boldsymbol{x}_a, \text{MelSpectrogram})$$
$$\boldsymbol{x}_{b'} = \boldsymbol{s} \odot \boldsymbol{x}_b + \boldsymbol{t}$$
$$\boldsymbol{f}^{-1}_{\text{coupling}}(\boldsymbol{x}) = \text{concat}(\boldsymbol{x}_a, \boldsymbol{x}_{b'})$$

Squeeze Operation



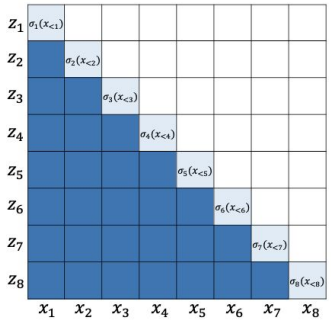
1x1 Invertible Convolution

- We want to permute **squeezed** channels
- Initialize weights as random **rotation** matrix

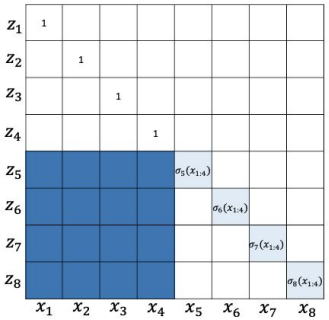


WaveFlow (WaveGlow + MAF)

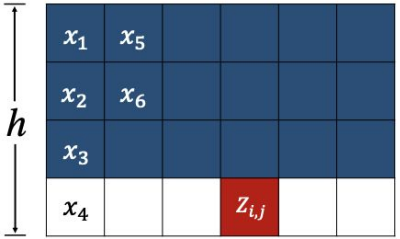
WaveFlow



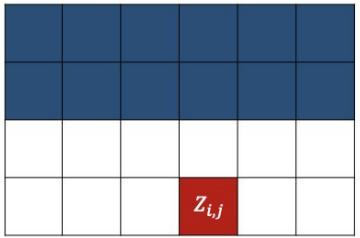
WaveGlow



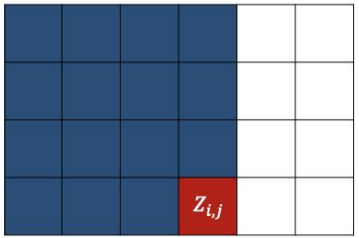
WaveFlow



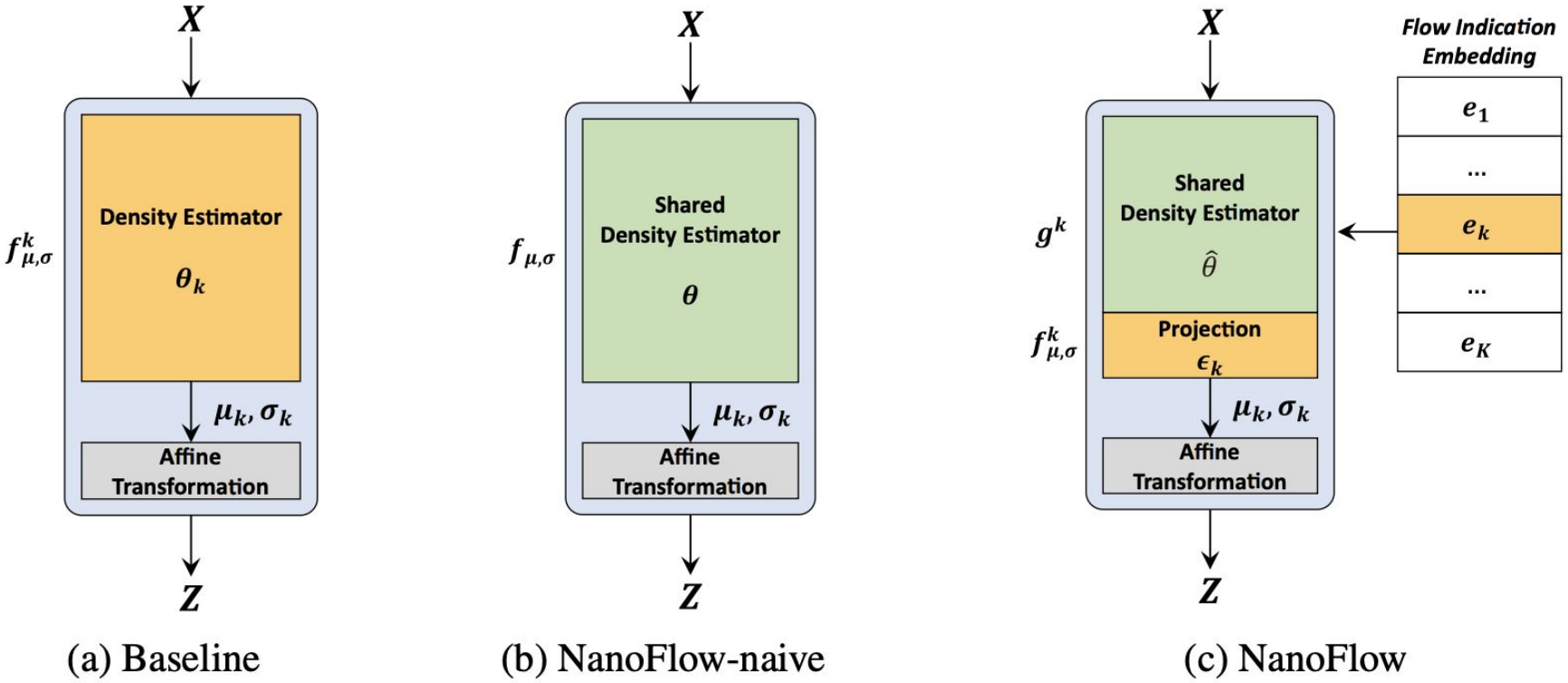
WaveGlow



“WaveNet”



NanoFlow

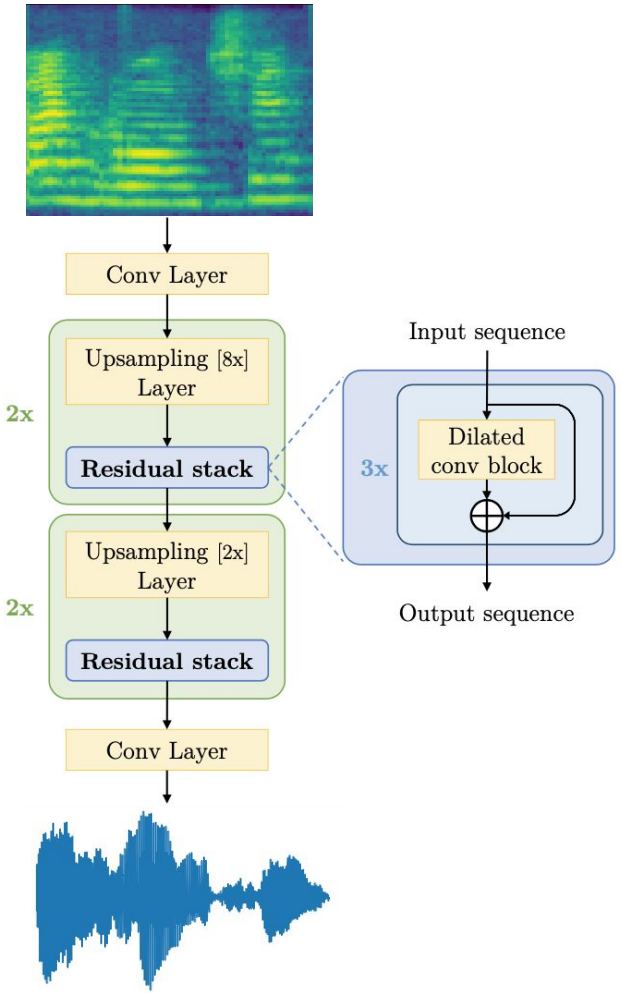


(a) Baseline (b) NanoFlow-naive (c) NanoFlow

What if… we just learn to map Mels to WAVs?

MelGAN

- Non-autoregressive
- Incredibly fast and don't require any kind of distillation



MelGAN

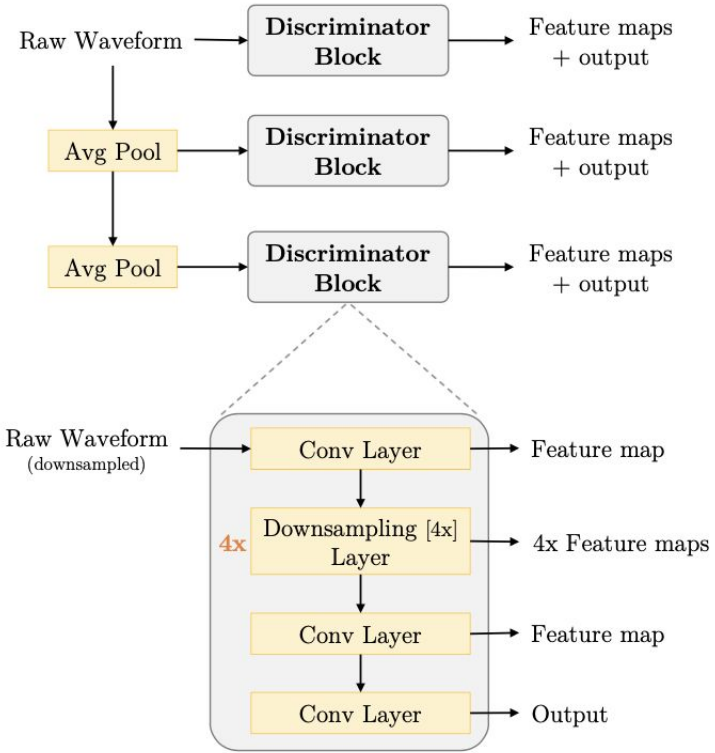
- Multiscale discriminator
- Feature Matching

$$\mathcal{L}_{\text{FM}}(G, D_k) = \mathbb{E}_{x,s\sim p_{\text{data}}}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D_k^{(i)}(x) - D_k^{(i)}(G(s))\right\|_1\right]$$

- Hinge Loss

$$\min_{D_k}\mathbb{E}_x[\min(0, 1 - D_k(x))] + \mathbb{E}_{s,z}[\min(0, 1 + D_k(G(s,z)))], \forall k = 1,2,3$$

$$\min_G \mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right]$$



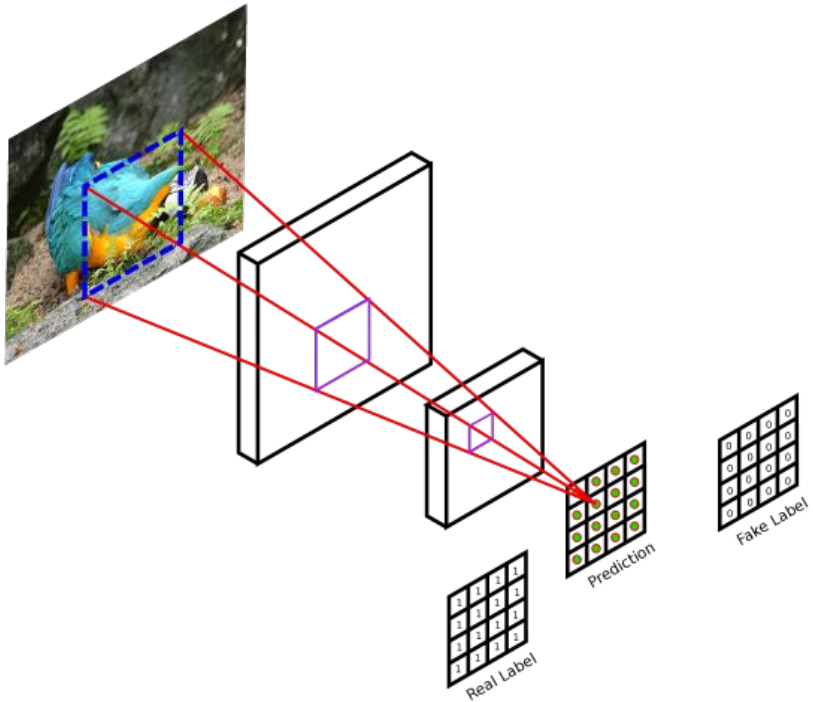
Weight Normalization

- Low-cost calculations
- Don’t store additional weight
- Don’t have train/test domain gap in statistics

$$\mathbf{w} = \frac{g}{\|\mathbf{v}\|}\mathbf{v}$$

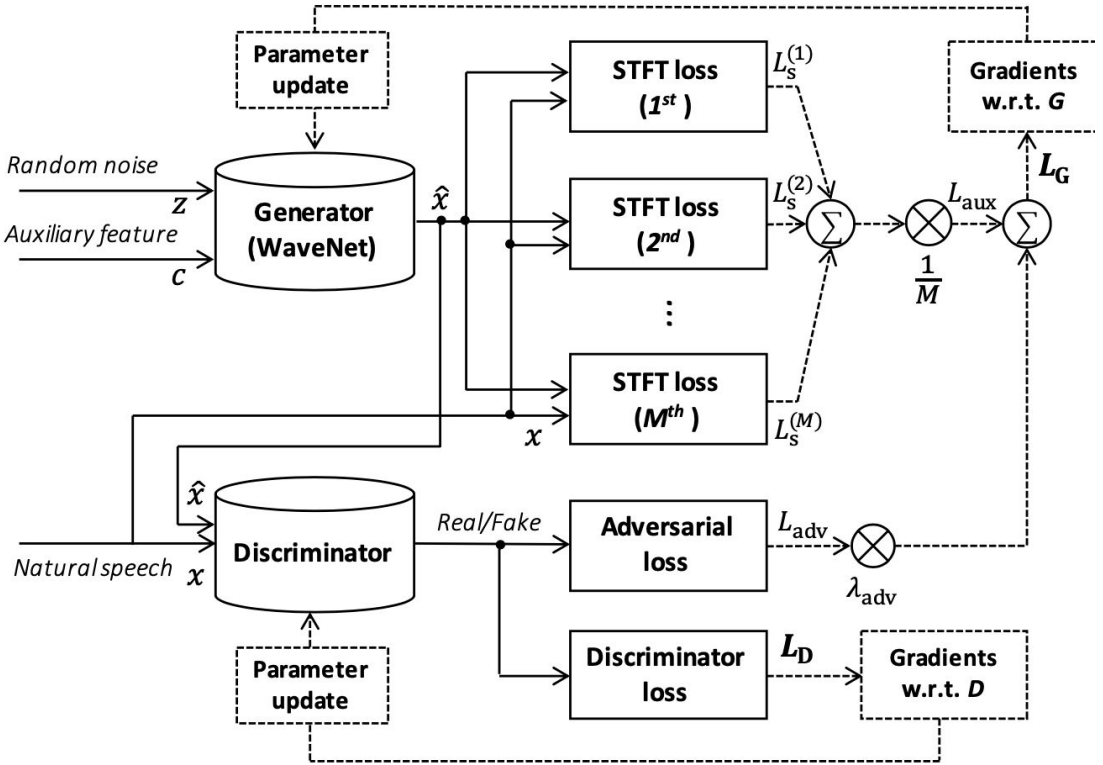
Markovian Discriminator

- Don’t classify entire audio sequences
- Classify random overlapped chunks



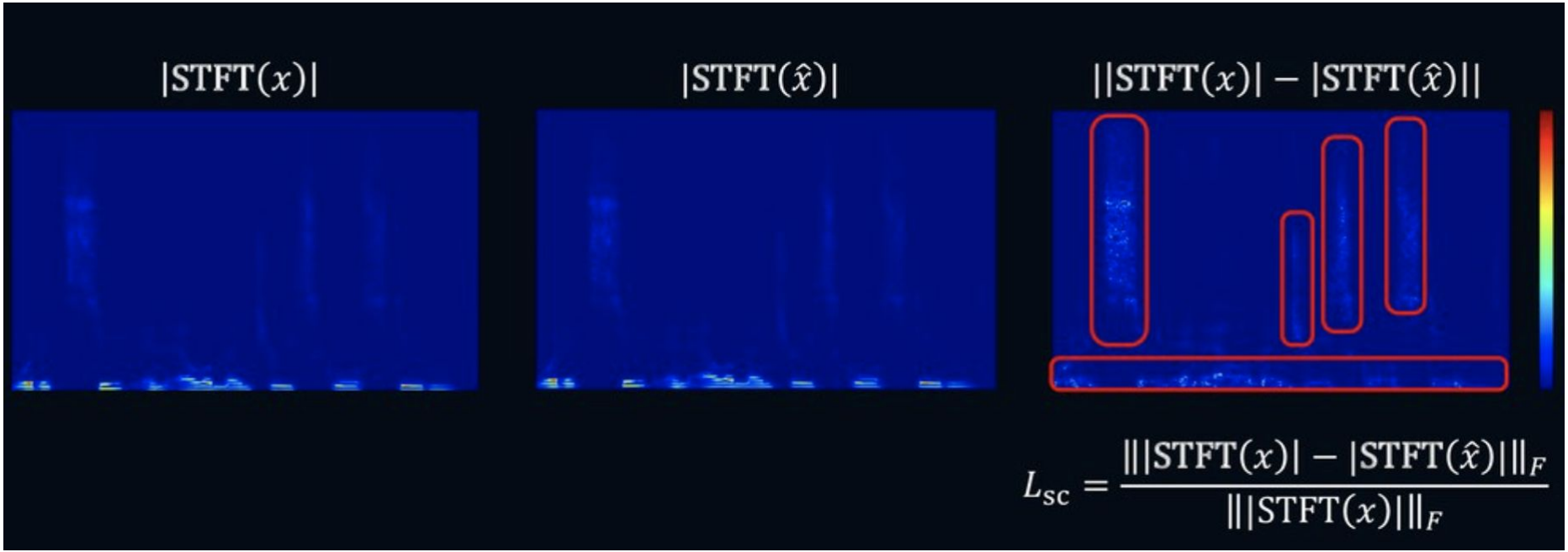
Parallel WaveGAN

- Similar to MelGAN
- Use WaveNet based Generator
- Additionally use **multi-STFT loss**
- **LSGAN** instead of Hinge



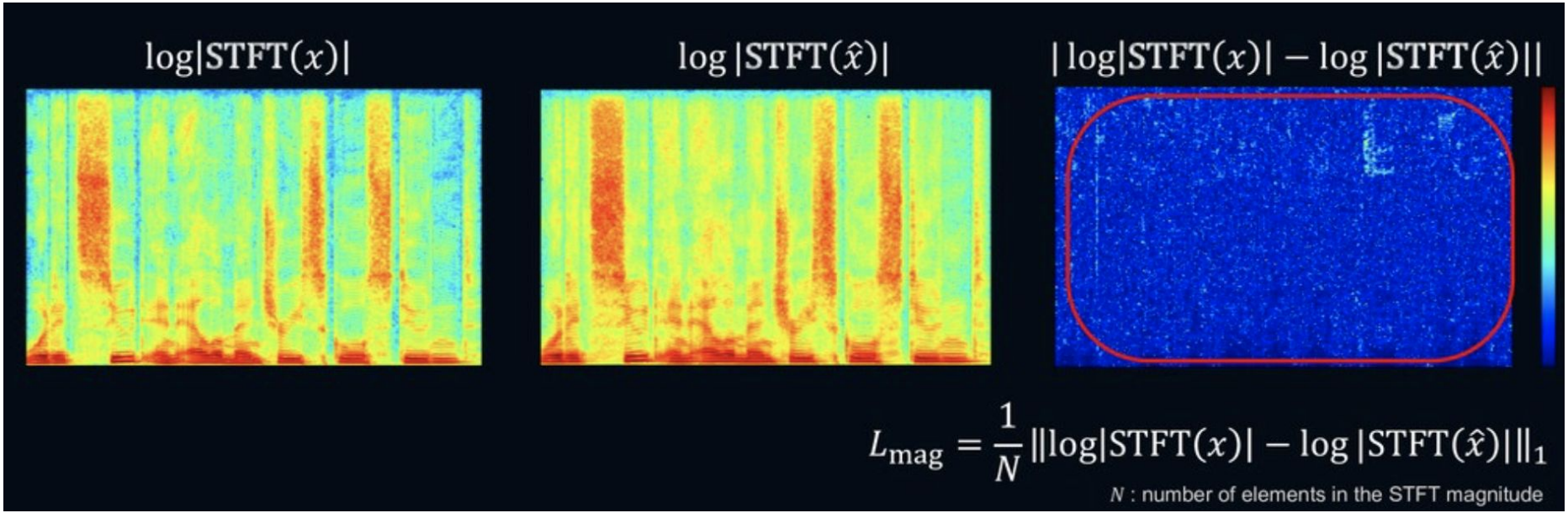
STFT Loss

Spectral Convergence part



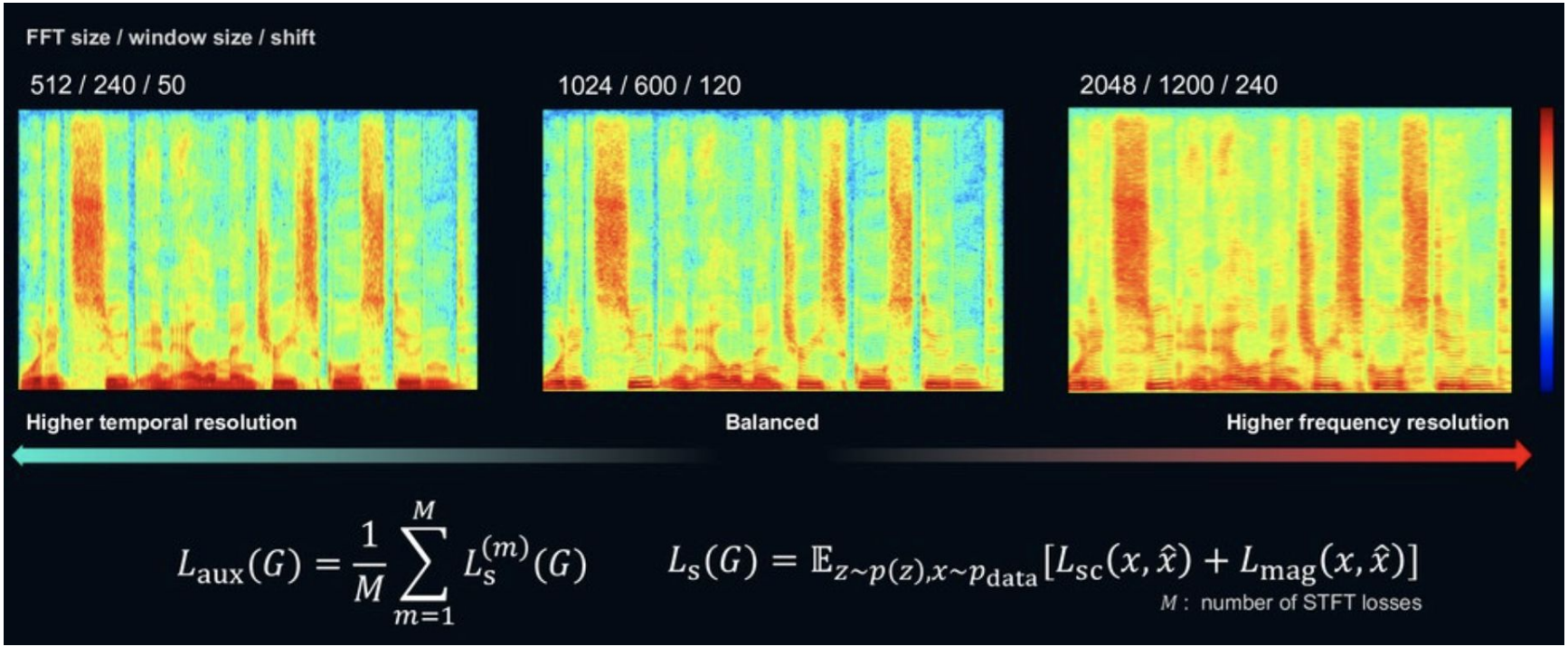
STFT Loss

Log scale STFT magnitude part



STFT Loss

Multi Resolution



What else?

- Use **torchaudio** and forget about **librosa!**
- Chinese vocoders aka **LPCNet**
- **WaveRNN**
- **Fast** WaveNet (with caches)
- More exotic **structural** losses as STFT
- More exotic discriminators
- Well designed generator/discriminators (HiFi GAN)
- (In my opinion) GANs **won** this war