Parallel WaveNet: Fast High-Fidelity Speech Synthesis

Kamlyk Erik 171

WaveNet vs. Parallel WaveNet

Figure 1: A second of generated speech.

WaveNet (2016)

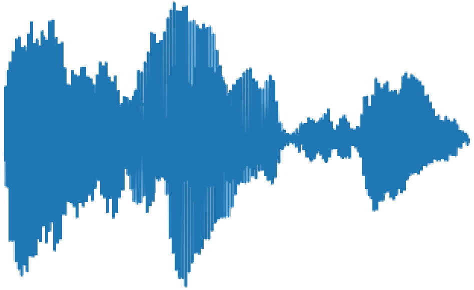
-high fidelity
-autoregressive
-slow inference

Parallel WaveNet (2017)

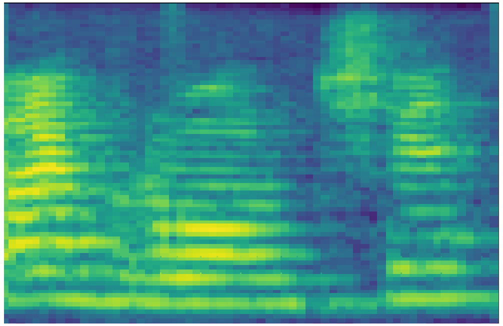
-distilled WaveNet
-non autoregressive
-fast inference
-very small quality drop

Overall pipeline

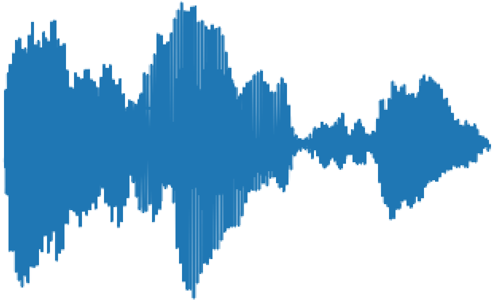
WaveNet: text → linguistic features →



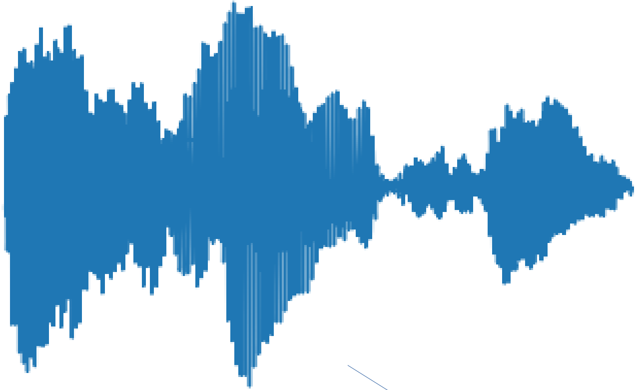
Tacotron2: text →

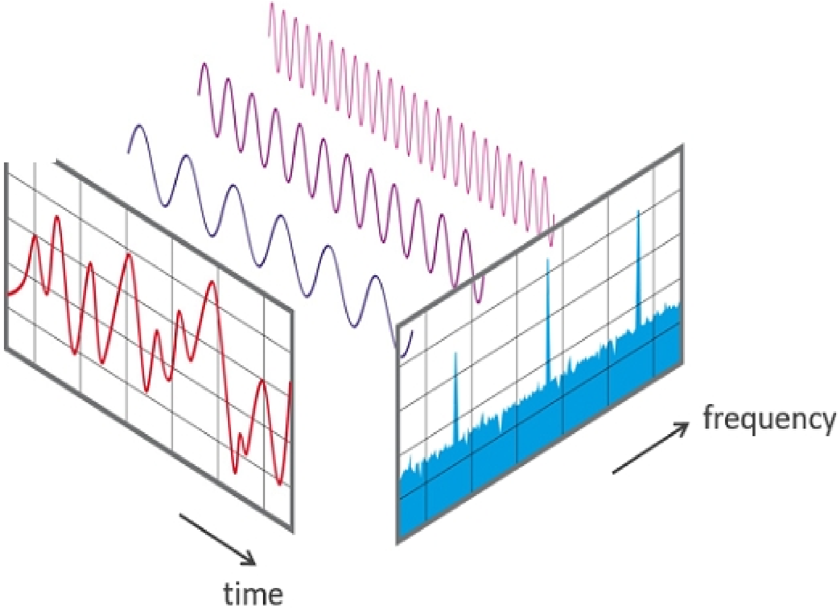


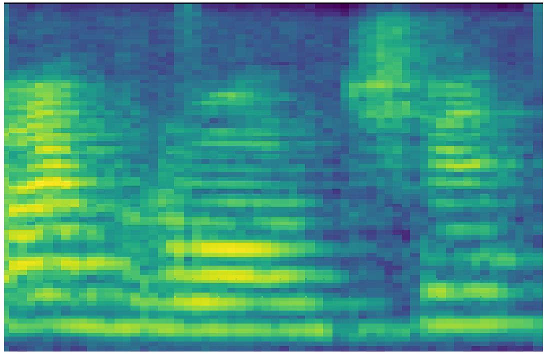
→



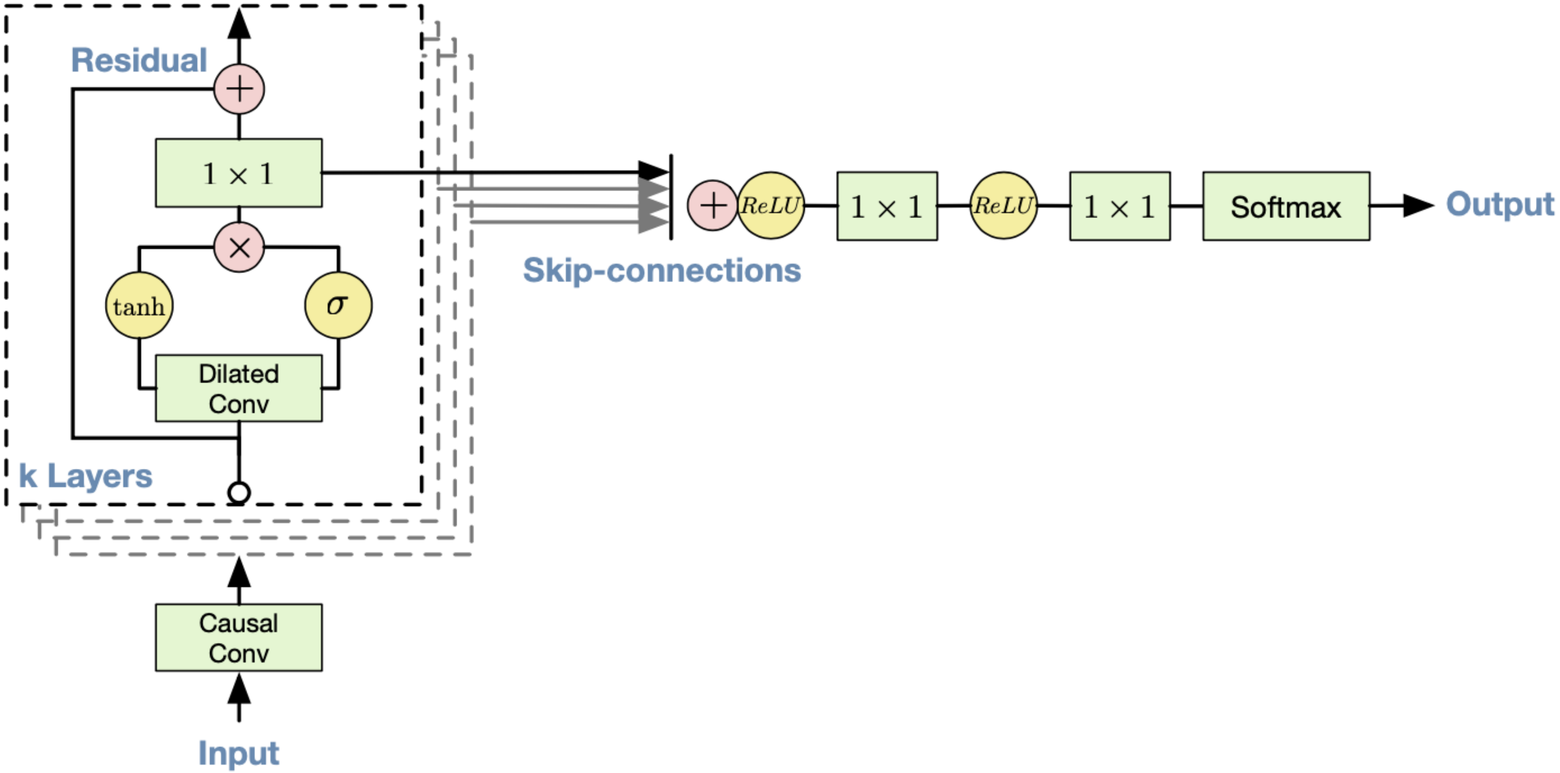
Mel Spectrogram





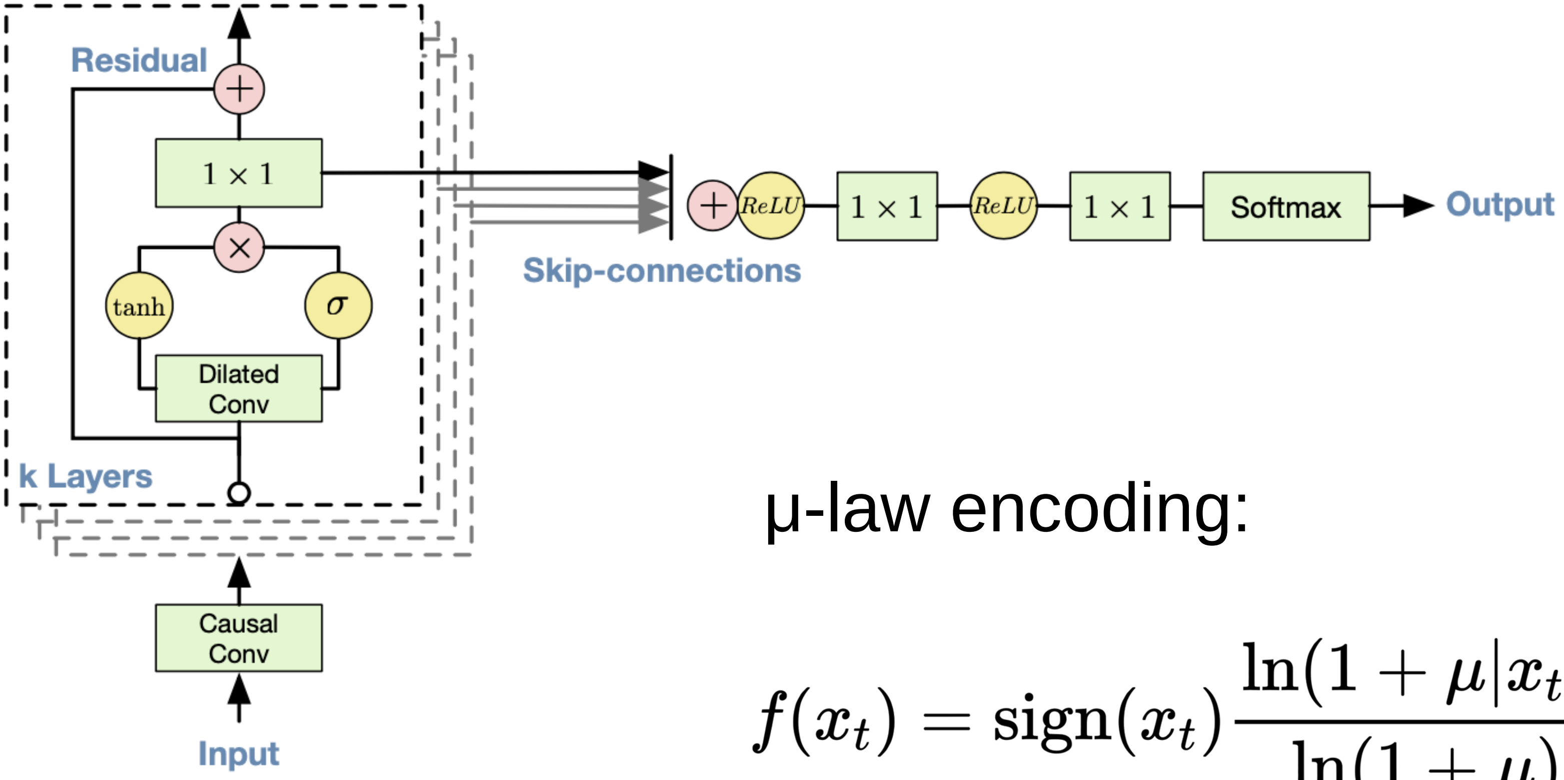


WaveNet



Residual
1×1
tanh
σ
Dilated Conv
k Layers
Causal Conv
Input
Skip-connections
$ReLU$
1×1
$ReLU$
1×1
Softmax
Output

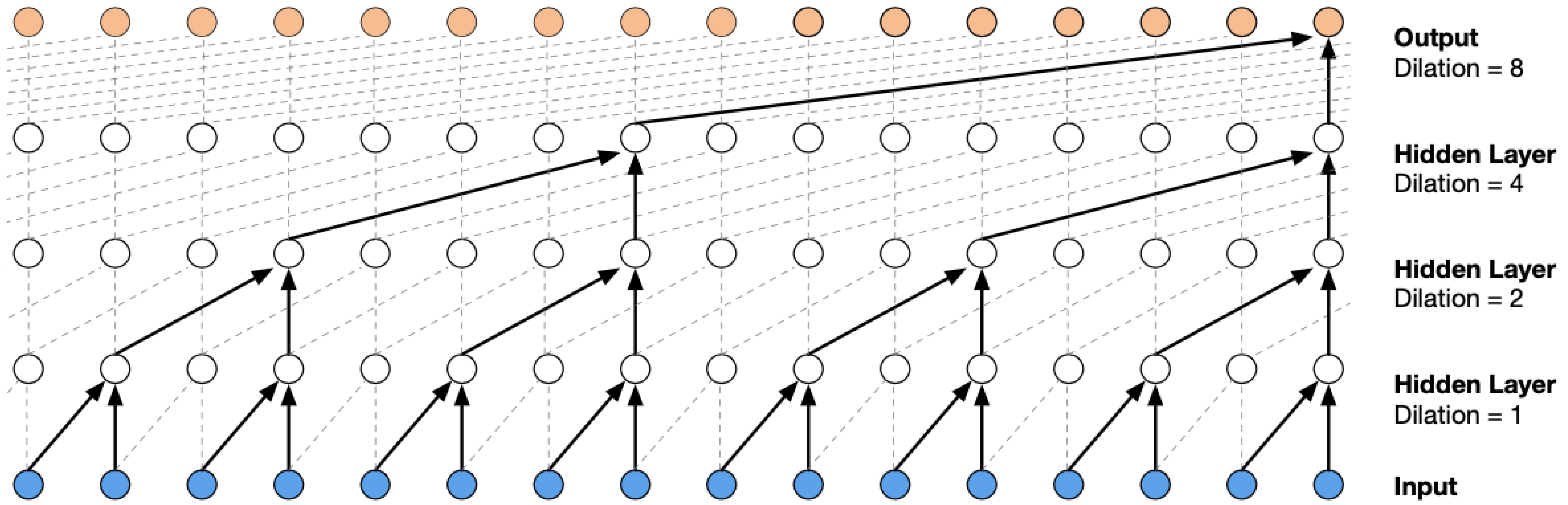
WaveNet



μ-law encoding:

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

WaveNet

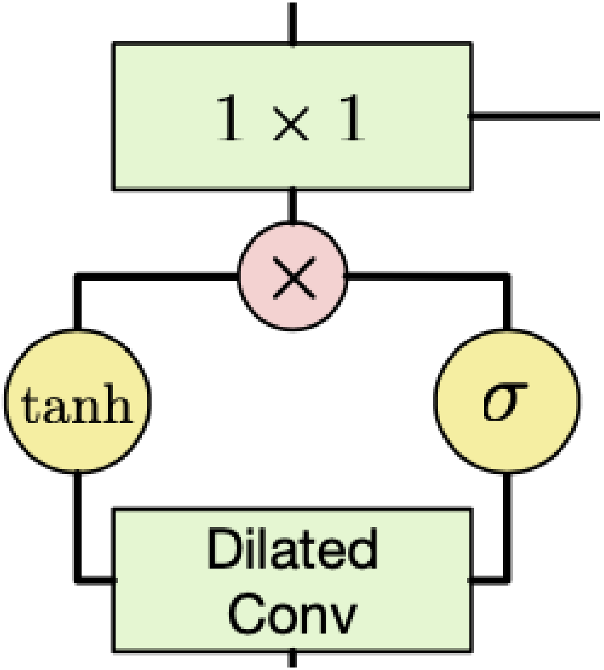


-causal convolutions (mask or padding)
-dilated convolutions

WaveNet

Gated layer:

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

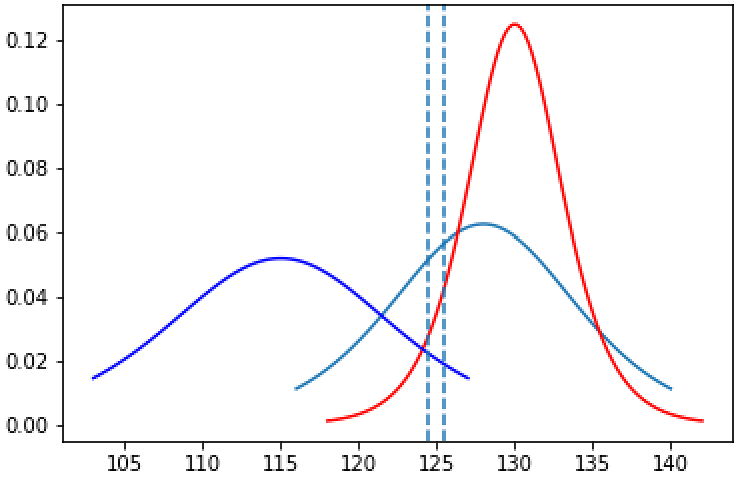


Conditioning on y:

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

High-fidelity WaveNet

1. 16-bit encoding (65536 values) instead of 8 bits (256 values).
2. Discretized mixture of logistics instead of softmax.
3. 24kHz instead of 16kHz (convolution filter size 2 → 3)



$$\nu \sim \sum_{i=1}^{K} \pi_i \text{logistic}(\mu_i, s_i)$$

$$P(x|\pi, \mu, s) = \sum_{i=1}^{K} \pi_i \left[\sigma((x + 0.5 - \mu_i)/s_i) - \sigma((x - 0.5 - \mu_i)/s_i)\right],$$

Normalizing flows

$$\boldsymbol{x} = f(\boldsymbol{z})$$

$$\log p_X(\boldsymbol{x}) = \log p_Z(\boldsymbol{z}) - \log \left| \frac{d\boldsymbol{x}}{d\boldsymbol{z}} \right|,$$

where $\left| \frac{d\boldsymbol{x}}{d\boldsymbol{z}} \right|$ is the determinant of the Jacobian of f

If f is invertible and its Jacobian determinant is easy to compute we can optimize maximum likelihood.

Composition of functions can be used for better quality:

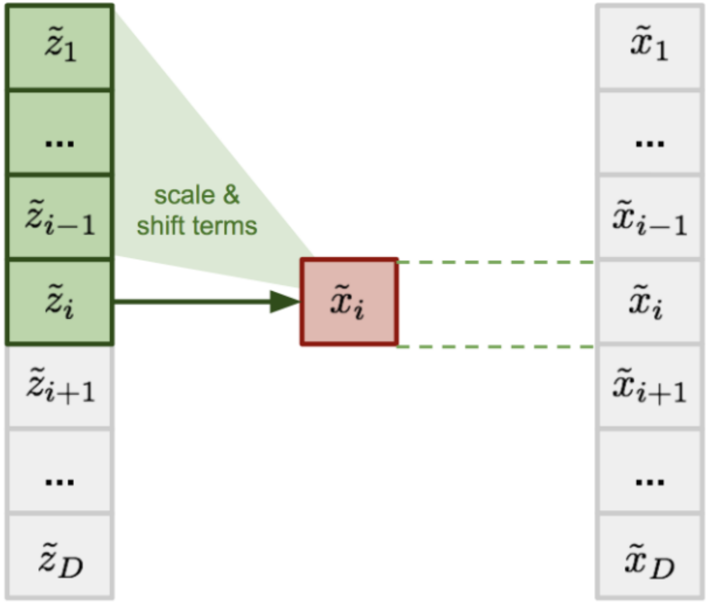
$$f_\theta = f^{(L)} \circ f^{(L-1)} \circ \ldots \circ f^{(1)}$$

Inverse Autoregressive Flows

$$x_t = f(\boldsymbol{z}_{\leq t})$$

$$\log \left| \frac{d\boldsymbol{x}}{d\boldsymbol{z}} \right| = \sum_t \log \frac{\partial f(\boldsymbol{z}_{\leq t})}{\partial z_t}$$

$$x_t = z_t \cdot s(\boldsymbol{z}_{<t}, \boldsymbol{\theta}) + \mu(\boldsymbol{z}_{<t}, \boldsymbol{\theta})$$



Parallel WaveNet

$$\boldsymbol{z} \sim \mathrm{Logistic}(0, I)$$

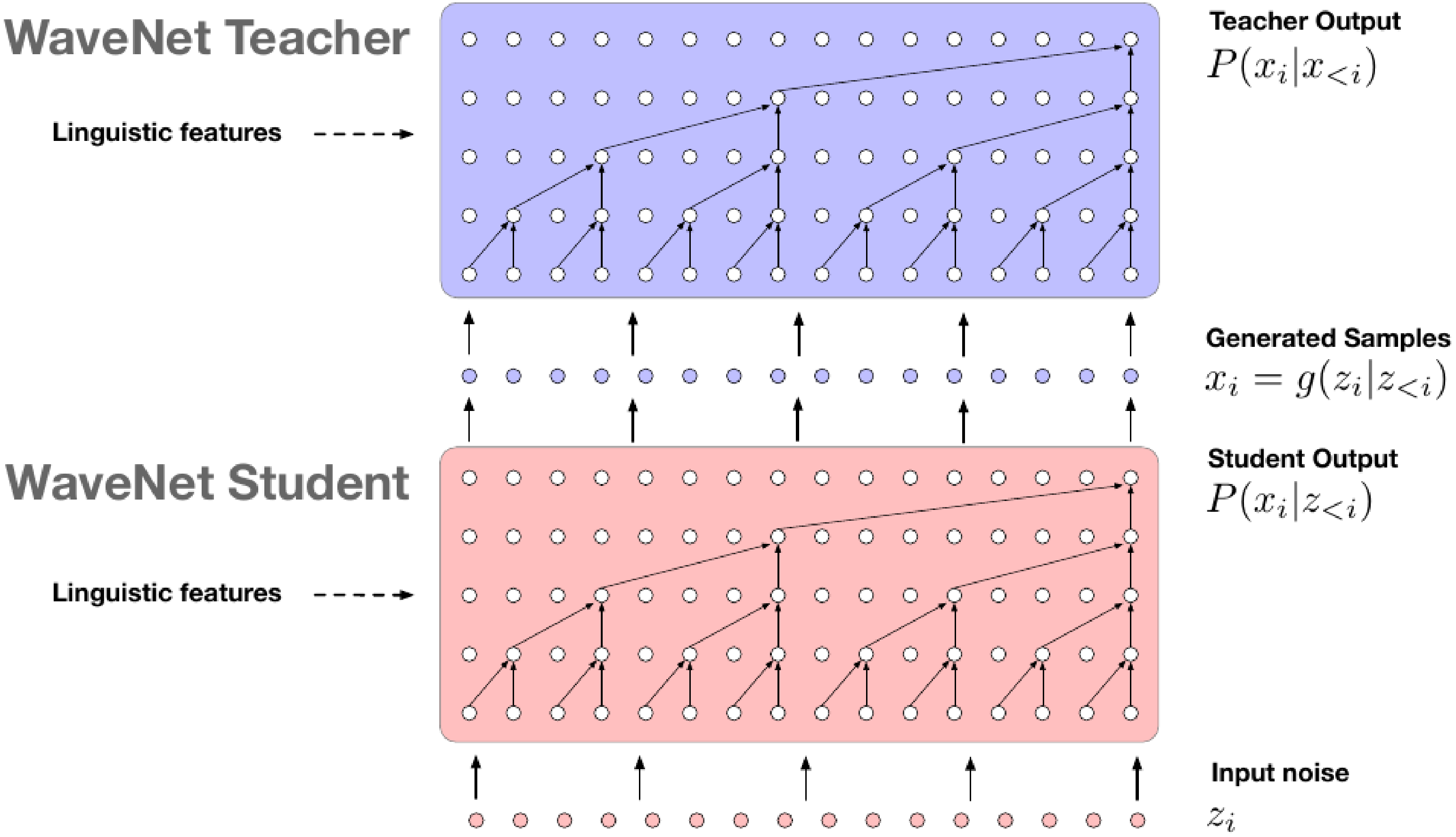
$$x_t = z_t \cdot s(\boldsymbol{z}_{<t}, \boldsymbol{\theta}) + \mu(\boldsymbol{z}_{<t}, \boldsymbol{\theta})$$

After N iterations:

$$\boldsymbol{\mu}_{\mathrm{tot}} = \sum_i^N \boldsymbol{\mu}^i \left(\prod_{j>i}^N \boldsymbol{s}^j \right)$$

$$\boldsymbol{s}_{\mathrm{tot}} = \prod_i^N \boldsymbol{s}_i$$

Training


WaveNet Teacher
Linguistic features
Teacher Output
$P(x_i|x_{<i})$
Generated Samples
$x_i = g(z_i|z_{<i})$
WaveNet Student
Linguistic features
Student Output
$P(x_i|z_{<i})$
Input noise
z_i

Loss function

$$D_{\mathrm{KL}}\left(P_S \| P_T\right) = H(P_S, P_T) - H(P_S)$$

$$\begin{aligned} H(P_S) &= \mathop{\mathbb{E}}_{z \sim L(0,1)}\left[\sum_{t=1}^{T} -\ln p_S(x_t|\boldsymbol{z}_{<t})\right] \\ &= \mathop{\mathbb{E}}_{z \sim L(0,1)}\left[\sum_{t=1}^{T} \ln s(\boldsymbol{z}_{<t}, \boldsymbol{\theta})\right] + 2T, \end{aligned}$$

$$\begin{aligned} H(P_S, P_T) &= \int_{\boldsymbol{x}} p_S(\boldsymbol{x}) \ln p_T(\boldsymbol{x}) \\ &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{p_S(\boldsymbol{x}_{<t})} H\Big(p_S(x_t|\boldsymbol{x}_{<t}), p_T(x_t|\boldsymbol{x}_{<t})\Big). \end{aligned}$$

Additional loss terms

1) Power loss: $$\|\phi(g(\boldsymbol{z}, \boldsymbol{c})) - \phi(\boldsymbol{y})\|^2$$

$$\phi(\boldsymbol{x}) = |\mathrm{STFT}(\boldsymbol{x})|^2$$

2) Perceptual loss

3) Contrastive loss:

$$D_{\mathrm{KL}}(P_S(\boldsymbol{c}_1) \| P_T(\boldsymbol{c}_1)) - \gamma D_{\mathrm{KL}}(P_S(\boldsymbol{c}_1) \| P_T \boldsymbol{c}_2))$$

Experiments: fidelity, speed

Method	Subjective 5-scale MOS
16kHz, 8-bit μ-law, 25h data:	
LSTM-RNN parametric [27]	3.67 ± 0.098
HMM-driven concatenative [27]	3.86 ± 0.137
WaveNet [27]	4.21 ± 0.081
24kHz, 16-bit linear PCM, 65h data:	
HMM-driven concatenative	4.19 ± 0.097
Autoregressive WaveNet	4.41 ± 0.069
Distilled WaveNet	4.41 ± 0.078

Table 1: Comparison of WaveNet distillation with the autoregressive teacher WaveNet, unit-selection (concatenative), and previous results from [27]. MOS stands for Mean Opinion Score.

WaveNet speed – 172 timesteps/second
Parallel WaveNet speed – over 500000 timesteps/second

Experiments: multispeaker

	Parametric	Concatenative	Distilled WaveNet
English speaker 1 (female - 65h data)	3.88	4.19	4.41
English speaker 2 (male - 21h data)	3.96	4.09	4.34
English speaker 3 (male - 10h data)	3.77	3.65	4.47
English speaker 4 (female - 9h data)	3.42	3.40	3.97
Japanese speaker (female - 28h data)	4.07	3.47	4.23

Table 2: Comparison of MOS scores on English and Japanese with multi-speaker distilled WaveNets. Note that some speakers sounded less appealing to people and always get lower MOS, however distilled parallel WaveNet always achieved significantly better results.

Ablation studies

Method	Preference Scores versus baseline concatenative system Win - Lose - Neutral
Losses used	
KL + Power	60% - 15% - 25%
KL + Power + Perceptual	66% - 10% - 24%
KL + Power + Perceptual + Contrastive (= default)	65% - 9% - 26%

Table 3: Performance with respect to different combinations of loss terms. We report preference comparison scores since their mean opinion scores tend to be very close and inconclusive.

Conclusions

1. WaveNet is an autoregressive model for generating waveforms from text; based on causal and dilated convolutions.
2. Parallel WaveNet is a model based on Inverse Autoregressive Flows. Model itself is not autoregressive, hence much faster generation.
3. Parallel WaveNet is slow to train, therefore it is trained using distillation from WaveNet.
4. The result is x1000 speedup without any drop in quality.

References

1. Parallel WaveNet: Fast High-Fidelity Speech Synthesis
https://arxiv.org/abs/1711.10433
2. WaveNet: A Generative Model for Raw Audio
https://arxiv.org/abs/1609.03499
3. Improving Variational Inference with Inverse Autoregressive Flow
4. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications
https://arxiv.org/abs/1701.05517

Вопросы:

1. Опишите архитектуру модели WaveNet (какие данные подаются на вход, как выглядит блок модели, какие свёртки используются, какая функция потерь).

2. Опишите процедуру обучения. Выпишите функцию потерь для Probability density distillation.

3. Выпишите формулу для contrastive loss. Зачем он используется?