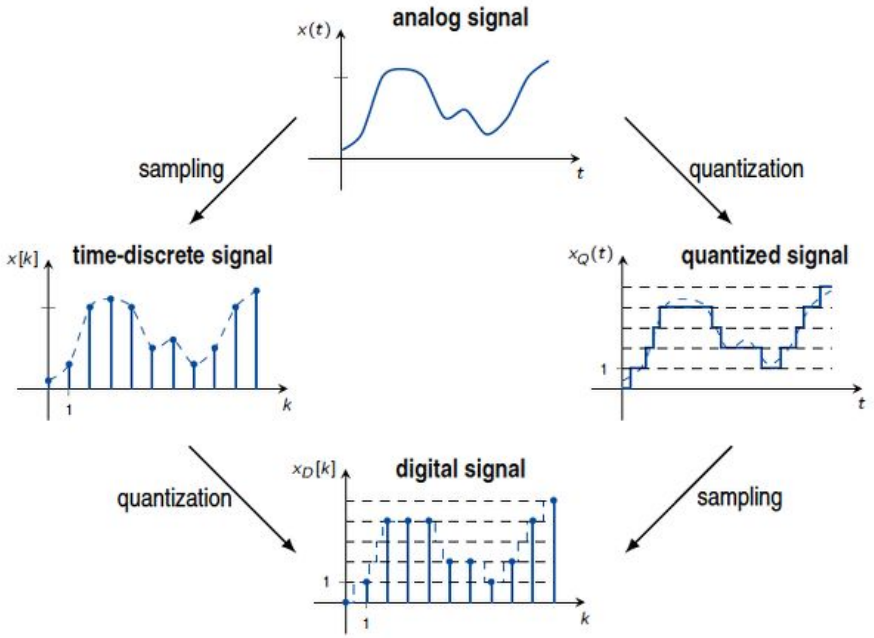
Speech Generation

Нейросетевые методы

Цифровое представление звука



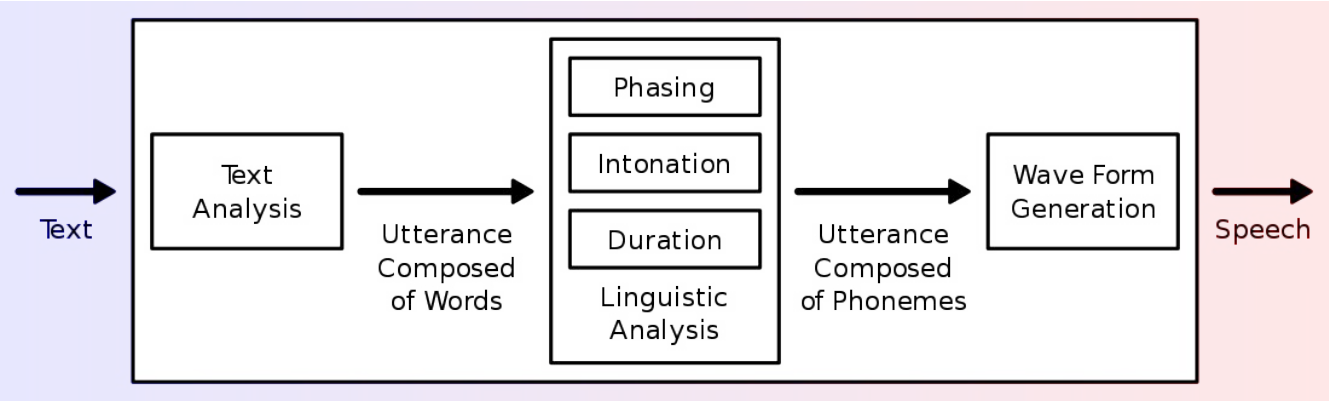
Sampling rate – количество сэмплов в секунду в цифровом представлении.

Bit depth – количество бит, кодирующих амплитуду сигнала в цифровом представлении.

Задача

Преобразование текста в речь, Text To Speech (TTS):

- Обработка текста.
- Синтез речи.



Основные методы синтеза

- Конкатенативный
- Статистический параметрический
- Нейросетевой

Нейросетевые модели

- WaveNet
- Tacotron
- DeepVoice
- Parallel WaveNet
- Tacotron 2

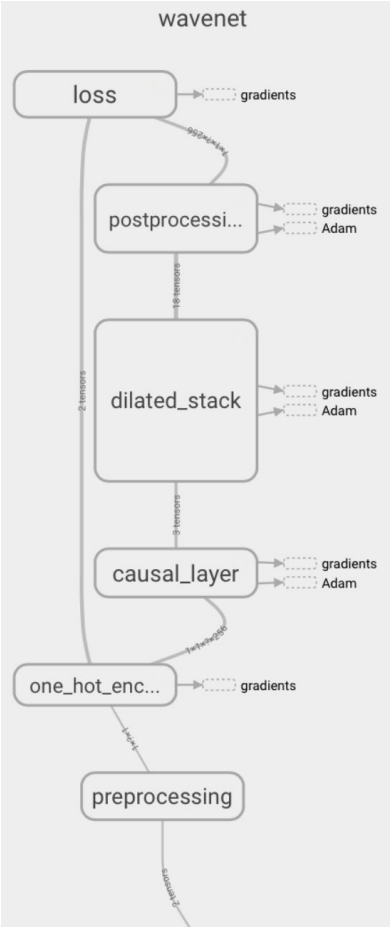
WaveNet

Генеративная нейронная сеть, сэмплирует из распределения звуковых сигналов. Решает задачу синтеза речи.

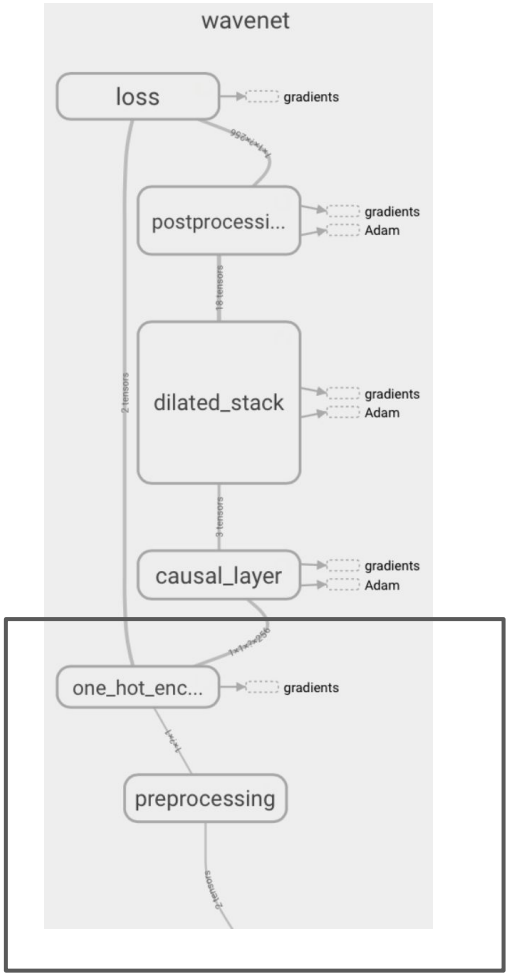
Вход: предыдущие предсказания (авторегрессия) + дополнительные параметры (лингвистические признаки, id говорящего, спектрограмма).

Выход: аудио сигнал.

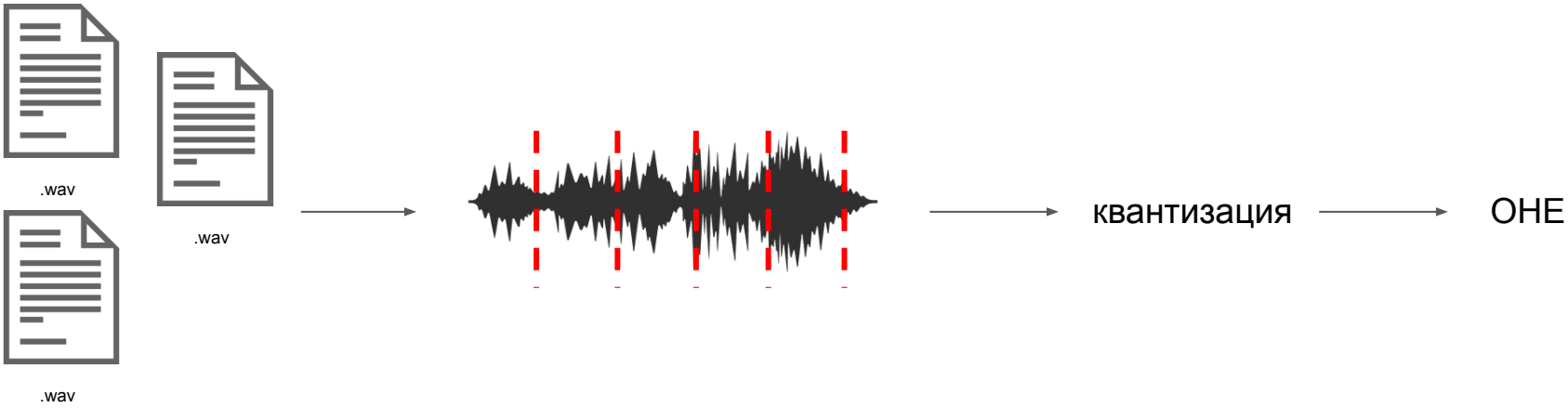
Обзор архитектуры WaveNet.



Обзор архитектуры WaveNet.

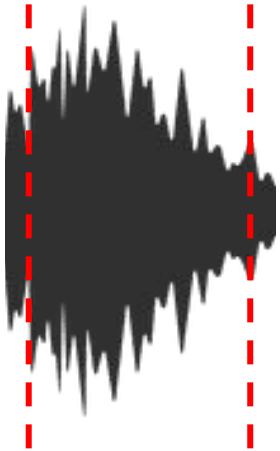


Preprocessing



Квантизация

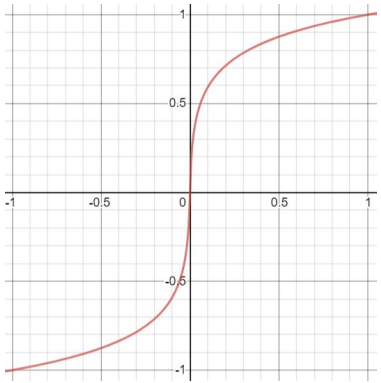
- Типичный bit depth сигнала от 16 до 32 битов. В конце сети softmax, следовательно придётся предсказывать от 2^16 категорий.



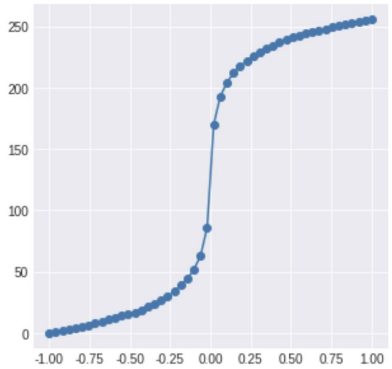
float32

μ-закон (μ=256)

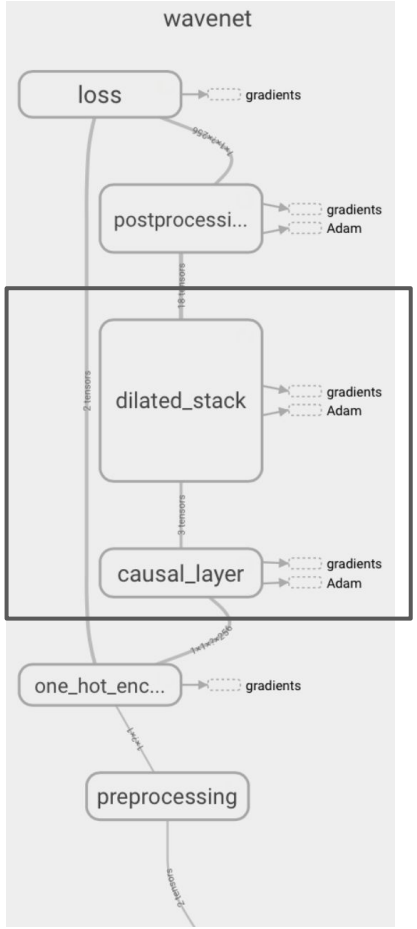
$$F(x) = \operatorname{sgn}(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)} \quad -1 \le x \le 1$$



квантизация в int8

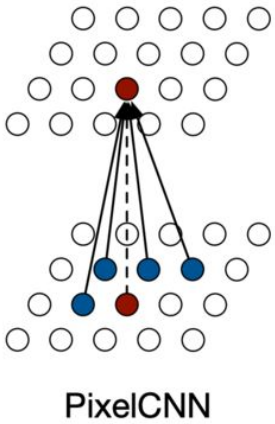


Обзор архитектуры WaveNet.

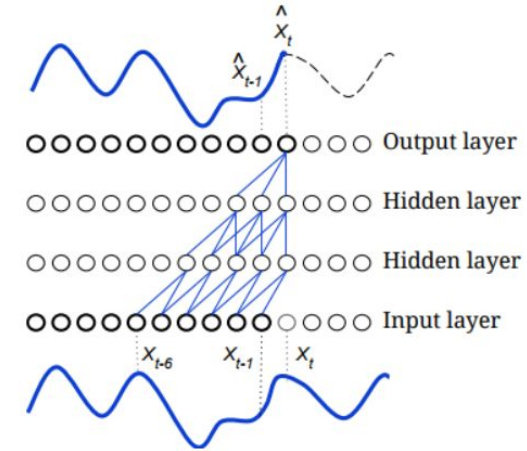


Causal convolution

Идея из PixelCNN. Гарантирует, что каждая единица сигнала зависит только от предыдущих.

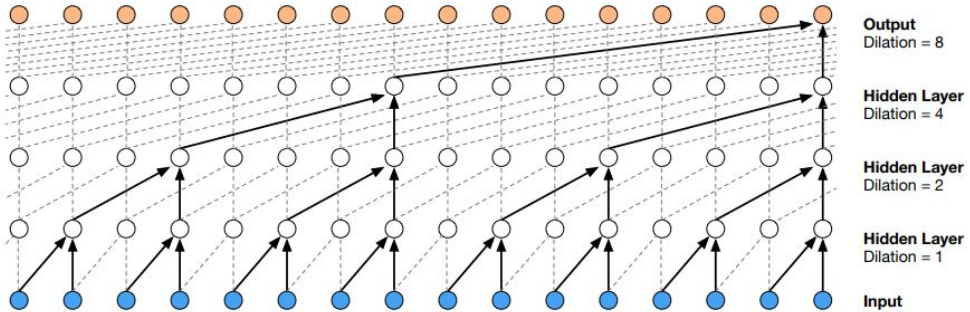


PixelCNN



Dilated convolution

Receptive field сети должен быть очень большой, поскольку 1 секунда звука ≈ 24000 предсказаний. Настраивается параметром dilation.



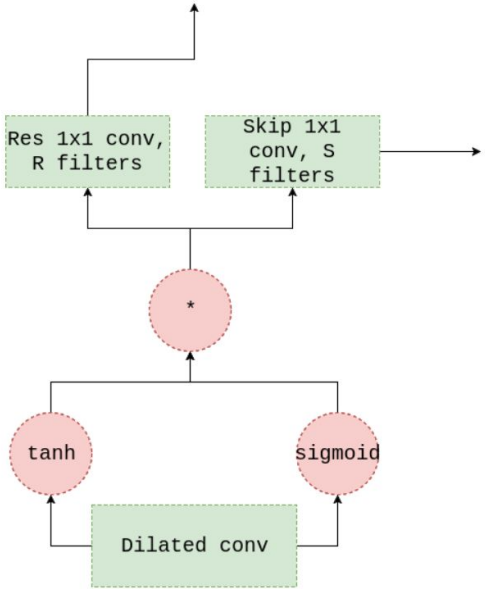
Функция активации

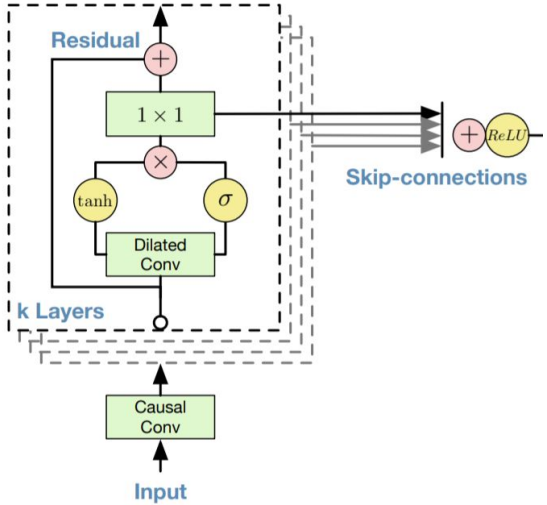
- Gated activation

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

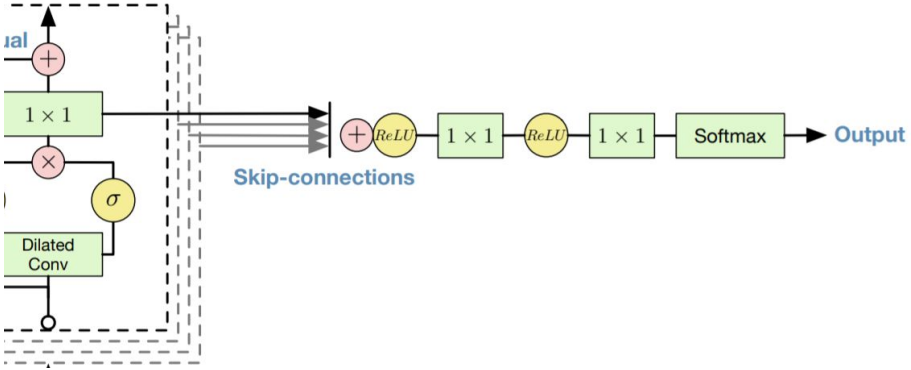
x – вход слоя.
$*$ – операция (dilated) causal свёртки.
$\odot$ – поэлементное умножение.
W_k – обучаемые фильтры свёртки на k-ом слое.

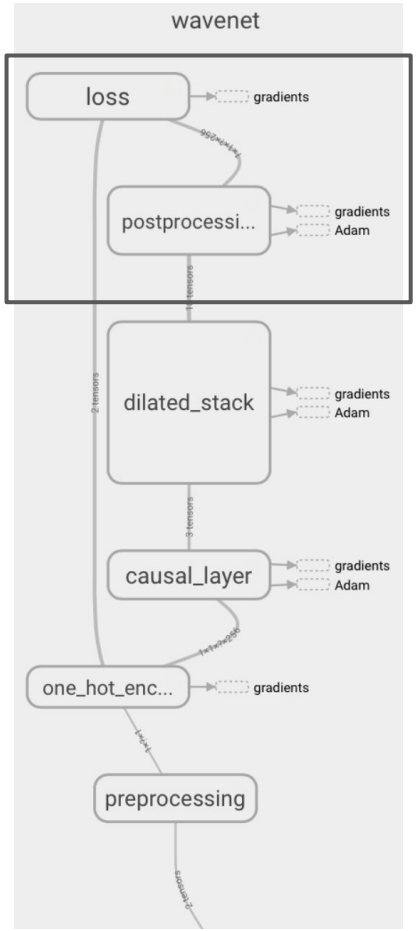
Skip и residual связи





Обзор архитектуры WaveNet.





Квантизация

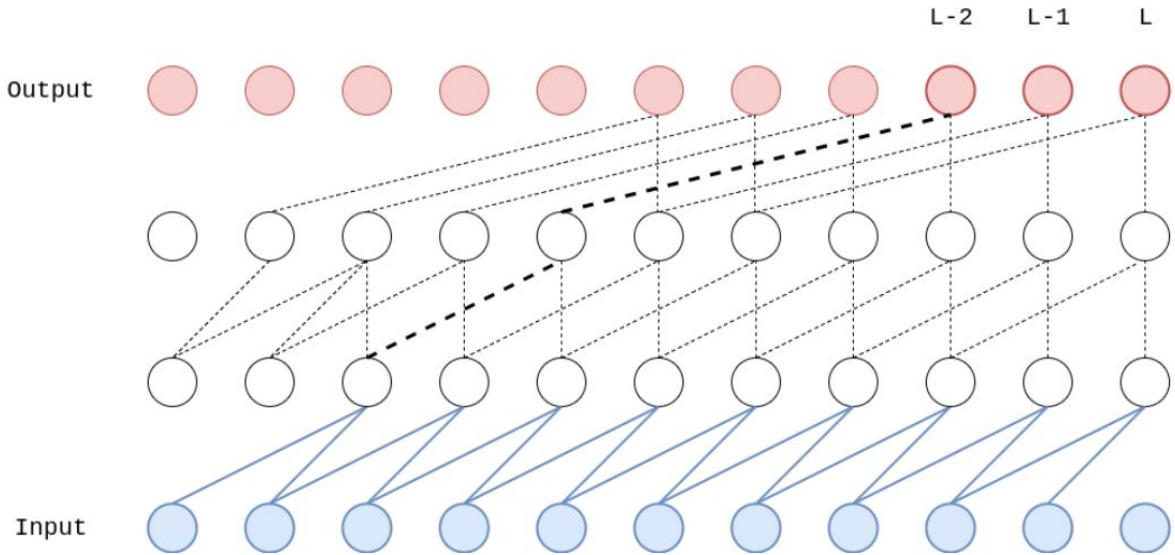
обратный μ-закон (μ=256)

softmax-256 → $F^{-1}(y) = \mathrm{sgn}(y)(1/\mu)((1+\mu)^{|y|} - 1) \quad -1 \le y \le 1$ →

.wav

Функция потерь

Кросс-энтропия между оригинальным сигналом и сгенерированным.



Global conditioning

Функция активации заменяется на:

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

V_k – обучаемые веса на k-ом слое.

h – вектор условия.

Local conditioning

Функция активации заменяется на:

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

V_k – обучаемые свёртки 1x1 на k-ом слое.

y – локальное условие (совпадает с **x** по размеру)

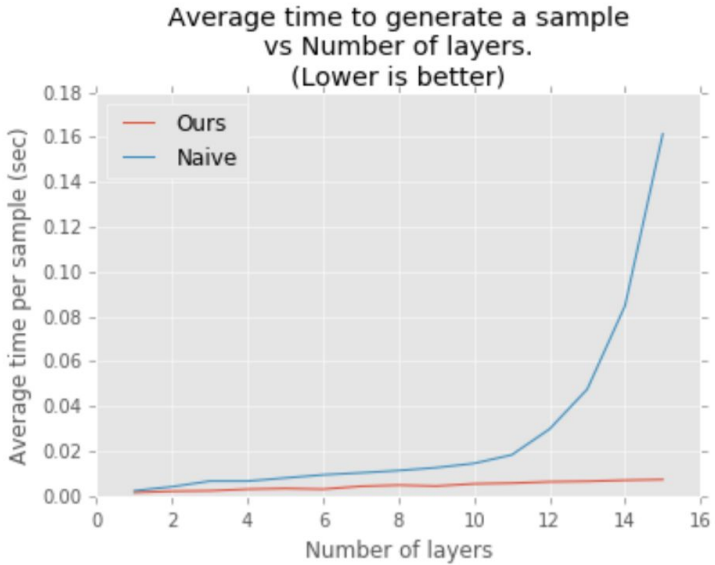
Результаты WaveNet

- Превзошла по натуральности параметрические и конкатенативные методы синтеза речи, используя лингвистические признаки.

	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Результаты WaveNet

- Далека от работы в реальном времени.



Tacotron

End-to-end решение для Text To Speech задачи.

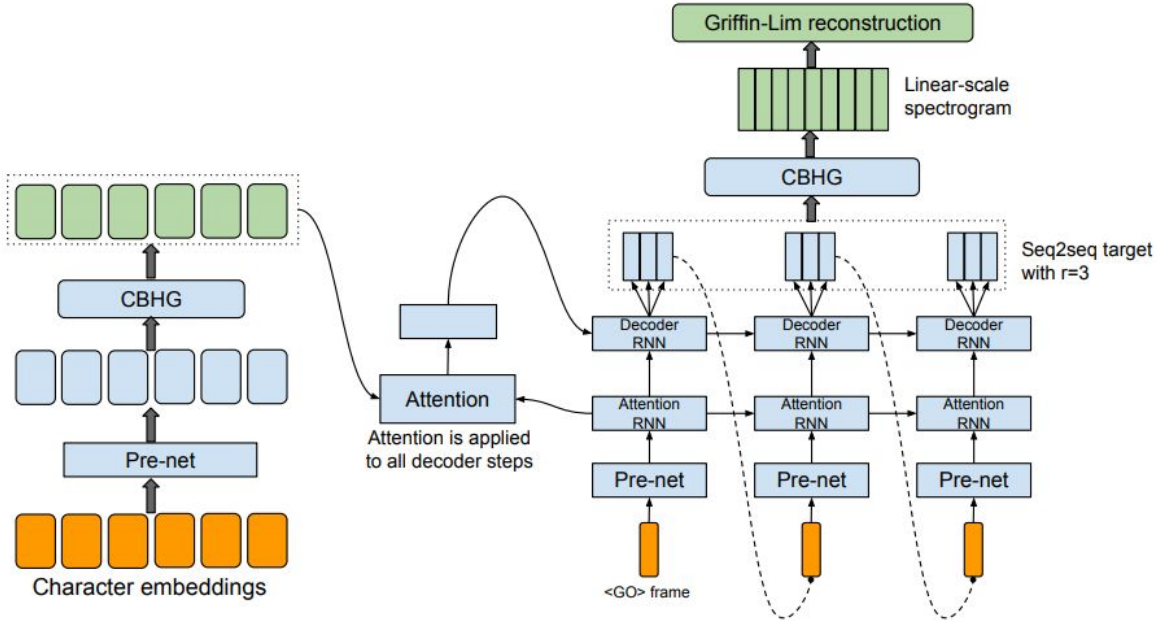
Вход: последовательность символов.

Выход: спектрограмма.

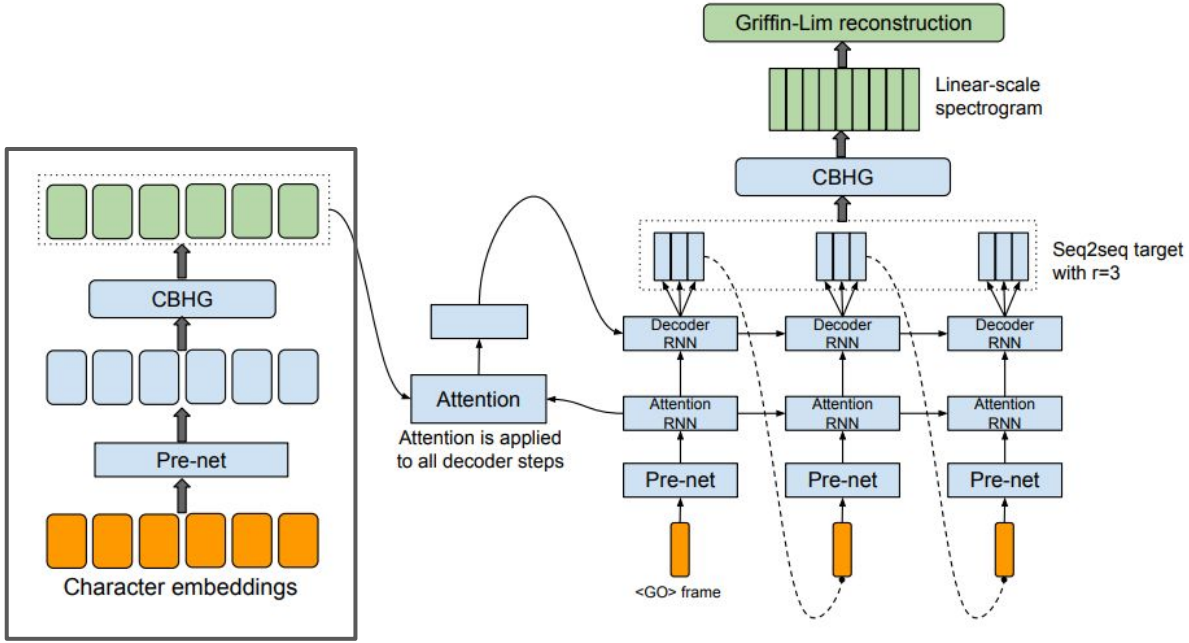
signal level for a given input. More...

*These authors really like tacos.
†These authors would prefer sushi.

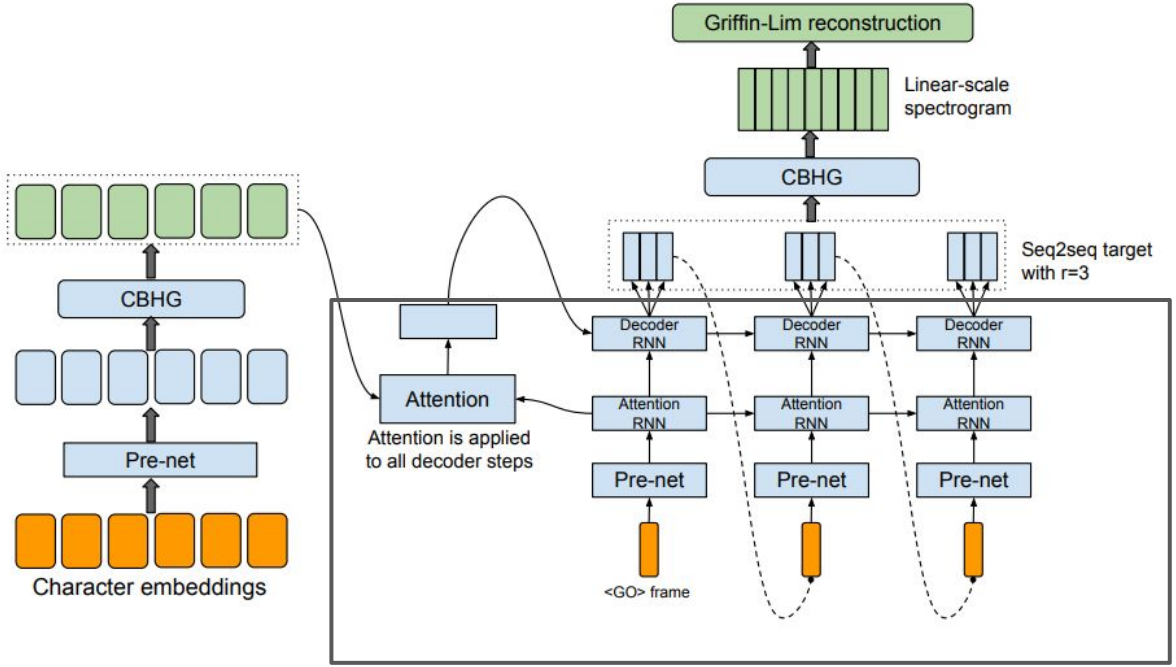
Обзор архитектуры Tacotron



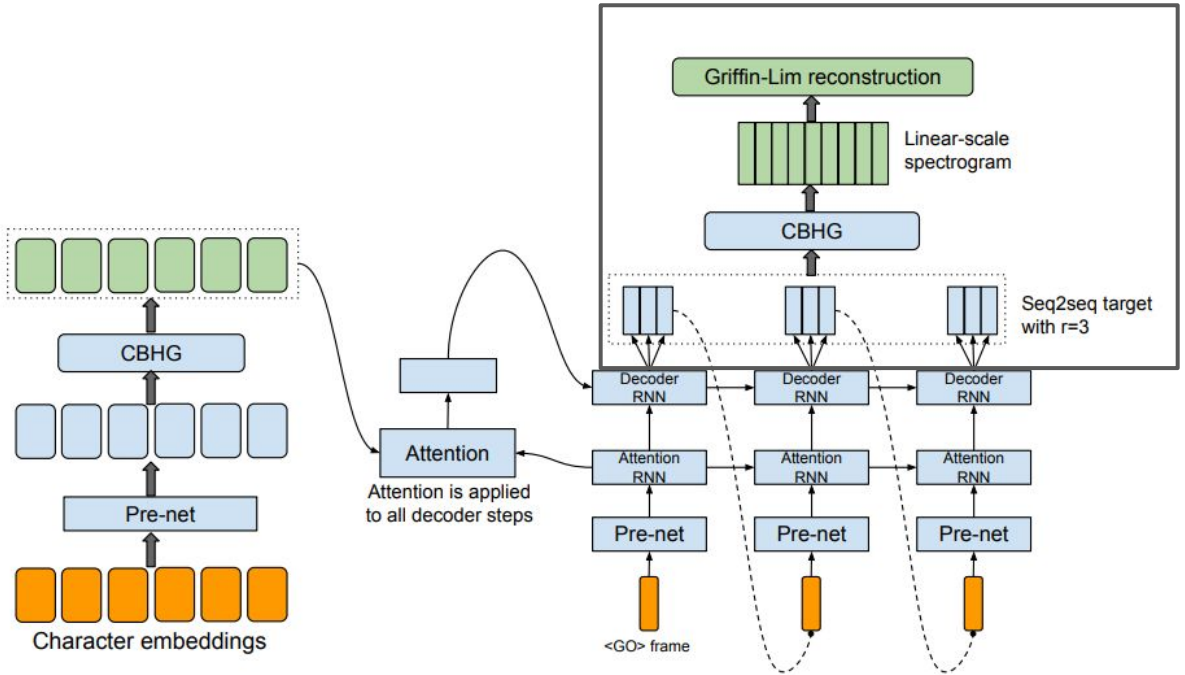
Обзор архитектуры Tacotron



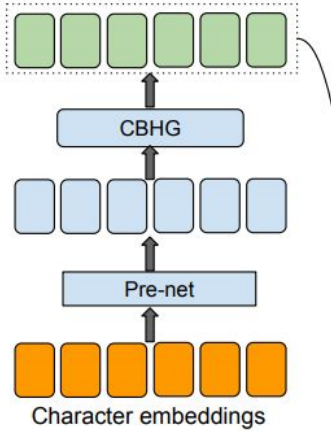
Обзор архитектуры Tacotron



Обзор архитектуры Tacotron



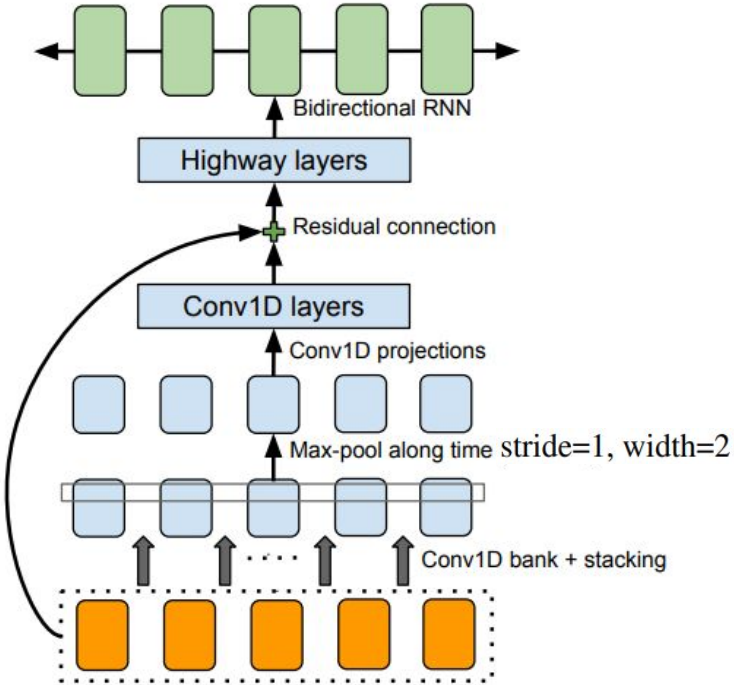
Encoder



Pre-net и embedding

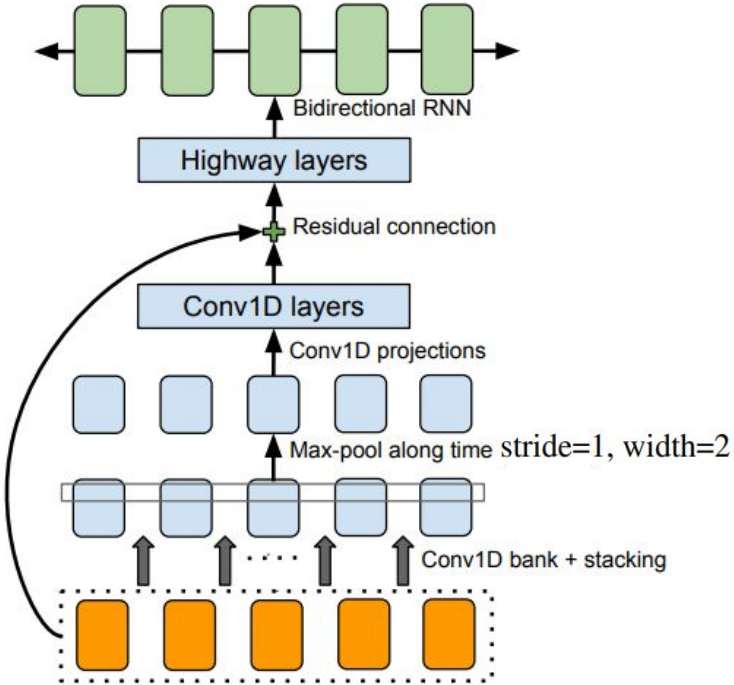
Encoder pre-net	FC-256-ReLU → Dropout(0.5) → FC-128-ReLU → Dropout(0.5)

CBHG модуль



Encoder CBHG	*Conv1D bank*: K=16, conv-k-128-ReLU *Max pooling*: stride=1, width=2 *Conv1D projections*: conv-3-128-ReLU $\rightarrow$ conv-3-128-Linear *Highway net*: 4 layers of FC-128-ReLU *Bidirectional GRU*: 128 cells

CBHG модуль



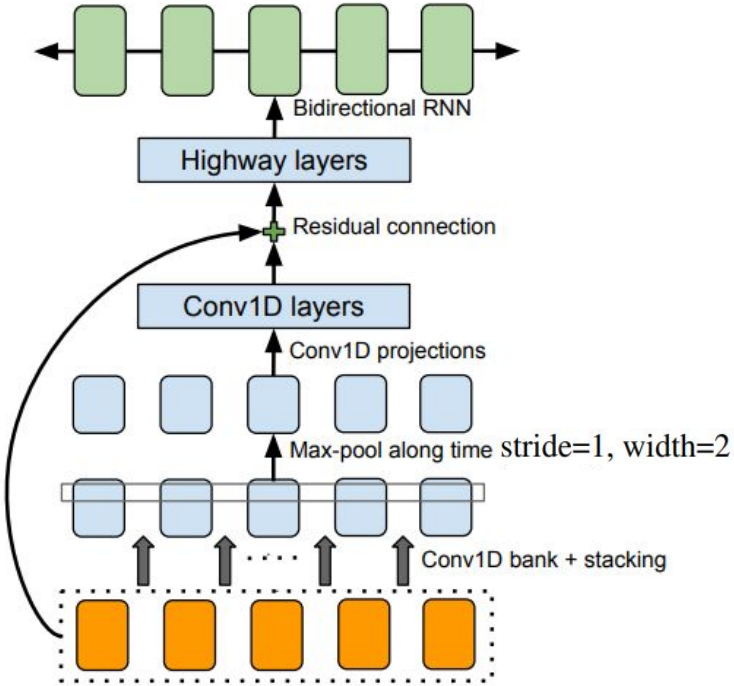
Highway layer

$$y = H(x, W_H)T(x, W_T) + x(\overline{1} - T(x, W_T))$$

$H(x, W_H)$ — обычный слой

$$T(x, W_T) = \sigma(W_t^T x + b_T)$$

CBHG модуль



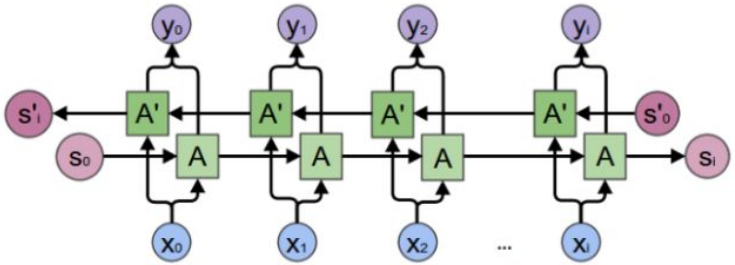
Highway layer

$$y = H(x, W_H)T(x, W_T) + x(\overline{1} - T(x, W_T))$$

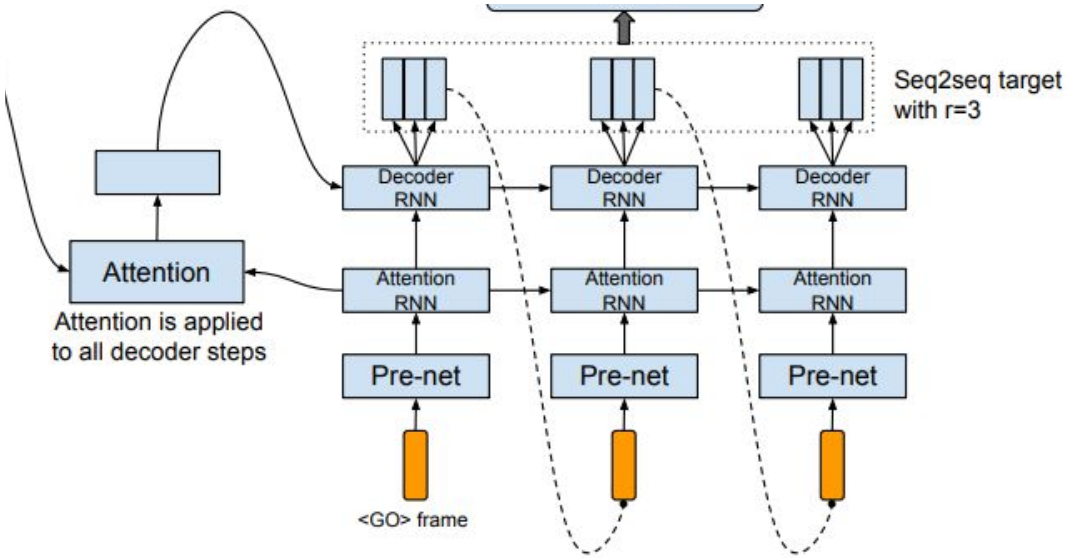
$H(x, W_H)$ — обычный слой

$$T(x, W_T) = \sigma(W_t^T x + b_T)$$

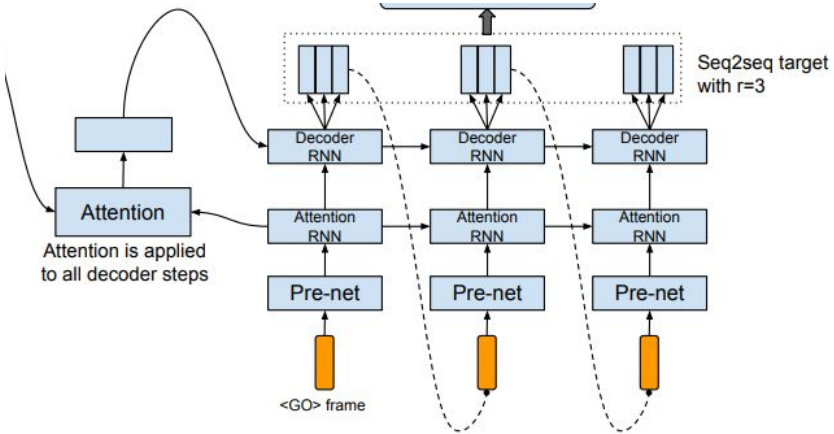
Bidirectional GRU



Decoder


Seq2seq target
with r=3
Decoder
RNN
Decoder
RNN
Decoder
RNN
Attention
Attention
RNN
Attention
RNN
Attention
RNN
Attention is applied
to all decoder steps
Pre-net
Pre-net
Pre-net
<GO> frame

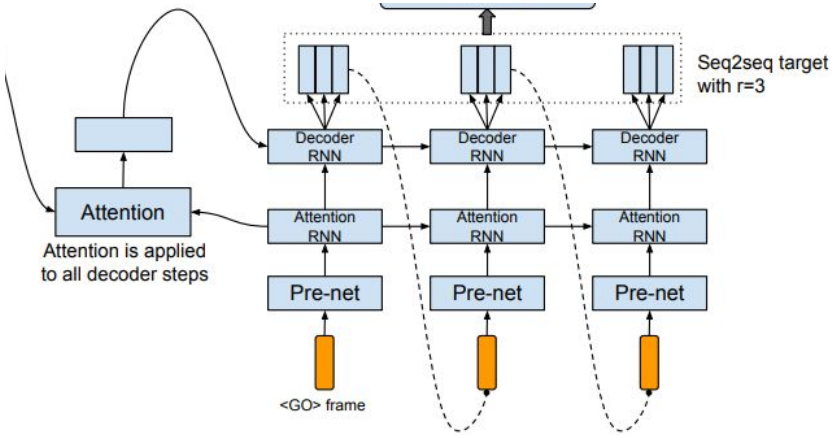
Decoder



Bahdanau Attention

Attention RNN	1-layer GRU (256 cells)

Decoder



Bahdanau Attention

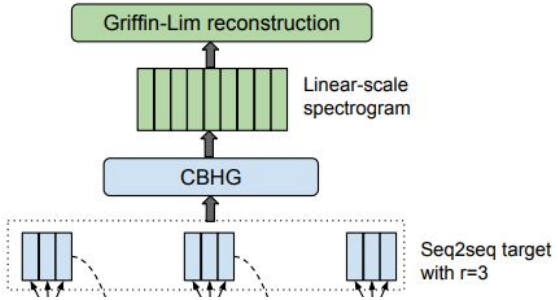
Decoder RNN

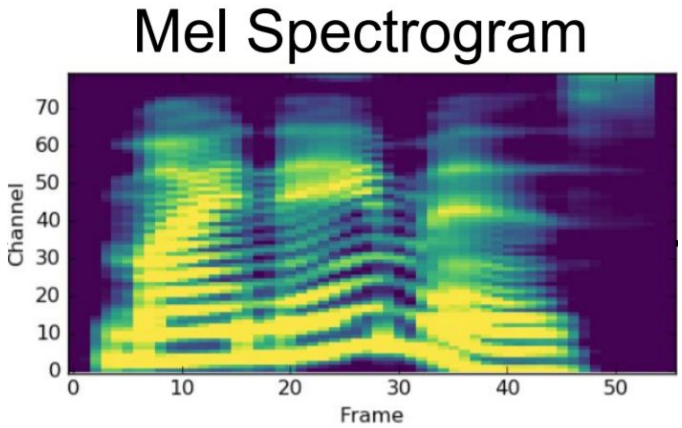
Attention RNN	1-layer GRU (256 cells)

Decoder RNN	2-layer residual GRU (256 cells)

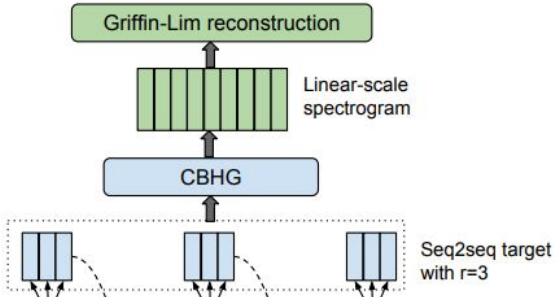
Post-processing

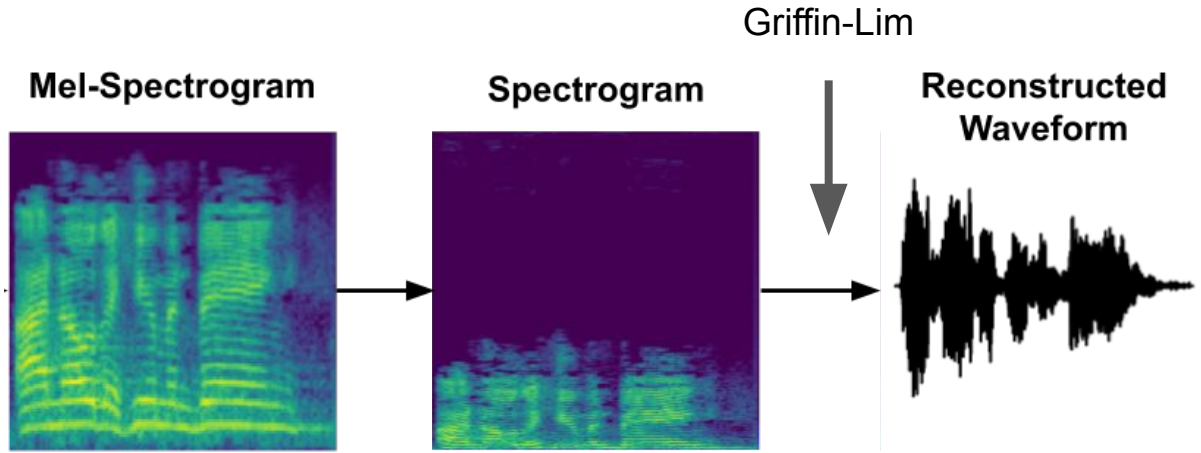
Выход decoder’a – мел-спектрограмма с небольшим числом диапазонов





Post-processing





Функция потерь

$$0.5 \cdot \frac{|mel_{predicted} - mel_{ground\ truth}|}{N} + 0.5 \cdot \frac{|linear_{predicted} - linear_{ground\ truth}|}{N}$$

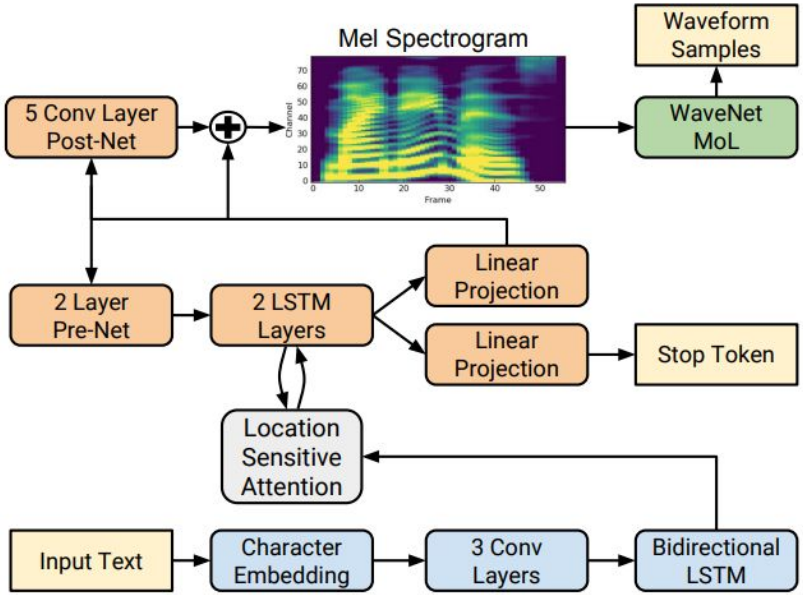
Результаты Tacotron

1 секунда аудио ≈ 0.22 секунды генерации

Table 2: 5-scale mean opinion score evaluation.

	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

WaveNet + Tacotron = Tacotron 2



System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Источники

- https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd#98f9
- https://medium.com/@kion.kim/wavenet-a-network-good-to-know-7caaae735435
- https://github.com/ibab/tensorflow-wavenet
- https://github.com/keithito/tacotron
- https://sergeiturukin.com/2017/03/02/wavenet.html
- https://arxiv.org/pdf/1611.09482.pdf
- https://arxiv.org/pdf/1712.05884.pdf
- https://arxiv.org/pdf/1703.10135.pdf