Tacotron 2:

NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

Общий поход к TTS

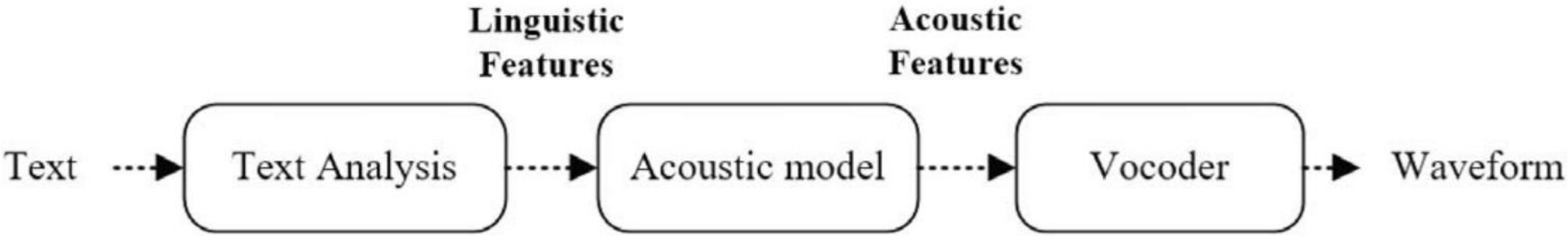


***Figure 1**: General Structure of TTS systems [4]*

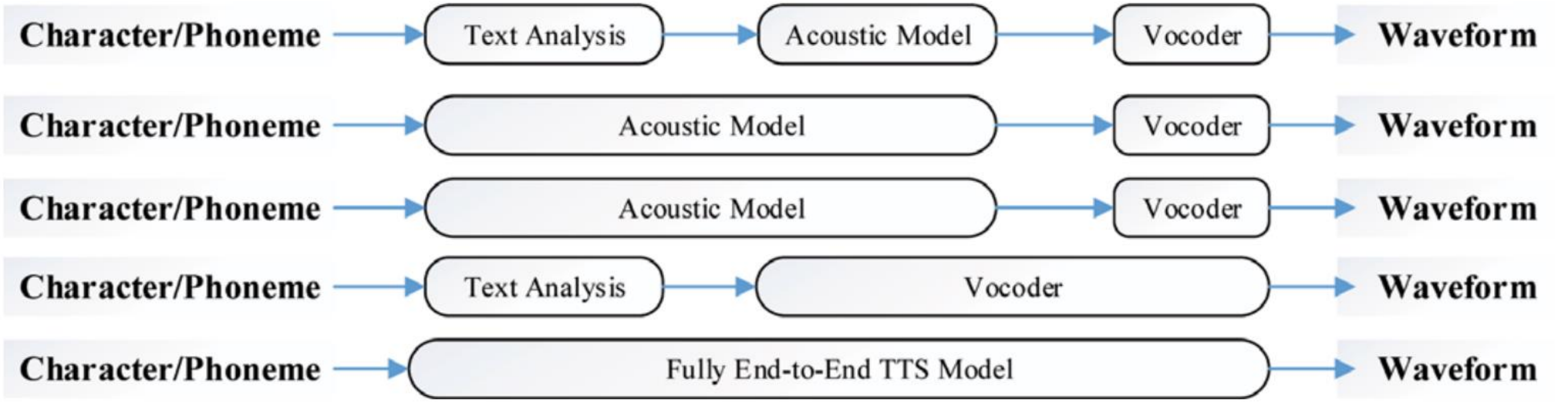


***Figure 2**: Different types of TTS model [4]*

Comparison of State-Of-The-Art TTS models

NAME	TYPE	BACKBONE ARCHITECTURE	INPUT	OUTPUT	VOCODER	DATASET (HOUR) (S-M*)	MOS
WaveNet	Autoregressive	PixelCNN	Linguistic Features	Wav	-	North American English (24.6) (S)	4.21±0.081
						Mandarin Chinese (34.8) (S)	4.08±0.085
Deep Voice	Autoregressive	CNN-based	Character / phoneme	Linguistic Features	WaveNet	Internal English Speech (20) (S) (Synthesized Duration and F0)	2.00±0.23
						Subset of the Blizzard 2013 (20.5)	2.67±0.37
Deep Voice 2	Autoregressive	CNN-based	Character / phoneme	Linguistic Features	WaveNet	VCTK (44) (M)	3.53±0.12
						Audiobooks (238) (M)	2.97±0.17
Deep Voice 3	Autoregressive	Fully CNN-based + attention + Seq2Seq	Character / phoneme	Acoustic Features	Griffin-Lim	VCTK (44) (M)	3.01 ±0.29
					WORLD		3.44±0.32
					WaveNet		-
CHAR2WAV	Autoregressive	Seq2Seq RNN	Character / Phoneme	Wav	SampleRNN	-	-
Tacotron	Autoregressive	RNN + Encoder-Decoder + Attention	Character / phoneme	Acoustic Features	Griffin-Lim	Internal North American English (24.6) (S)	3.82±0.085
Tacotron 2	Autoregressive	RNN + Encoder-Decoder + Attention	Character / phoneme	Acoustic Features	WaveNet	Internal North American English (24.6) (S)	4.5±0.06
Transformer TTS	Autoregressive	Transformer-based	Character / phoneme	Acoustic Features	WaveNet	Internal US English (25) (S)	4.39
ClariNet	Autoregressive	CNN-based	Character / phoneme	Wav	WaveNet	Internal English speech dataset (20)	4:15 ± 0:25

Tacotron

- Сходства со Tacotron 2:
  - RNN
  - Encoder and Decoder + Attention
  - Спектрограмма как промежуточное представление
- Отличия:
  - Более сложная архитектура
  - Линейная спектрограмма
  - GLA в качестве вокодера

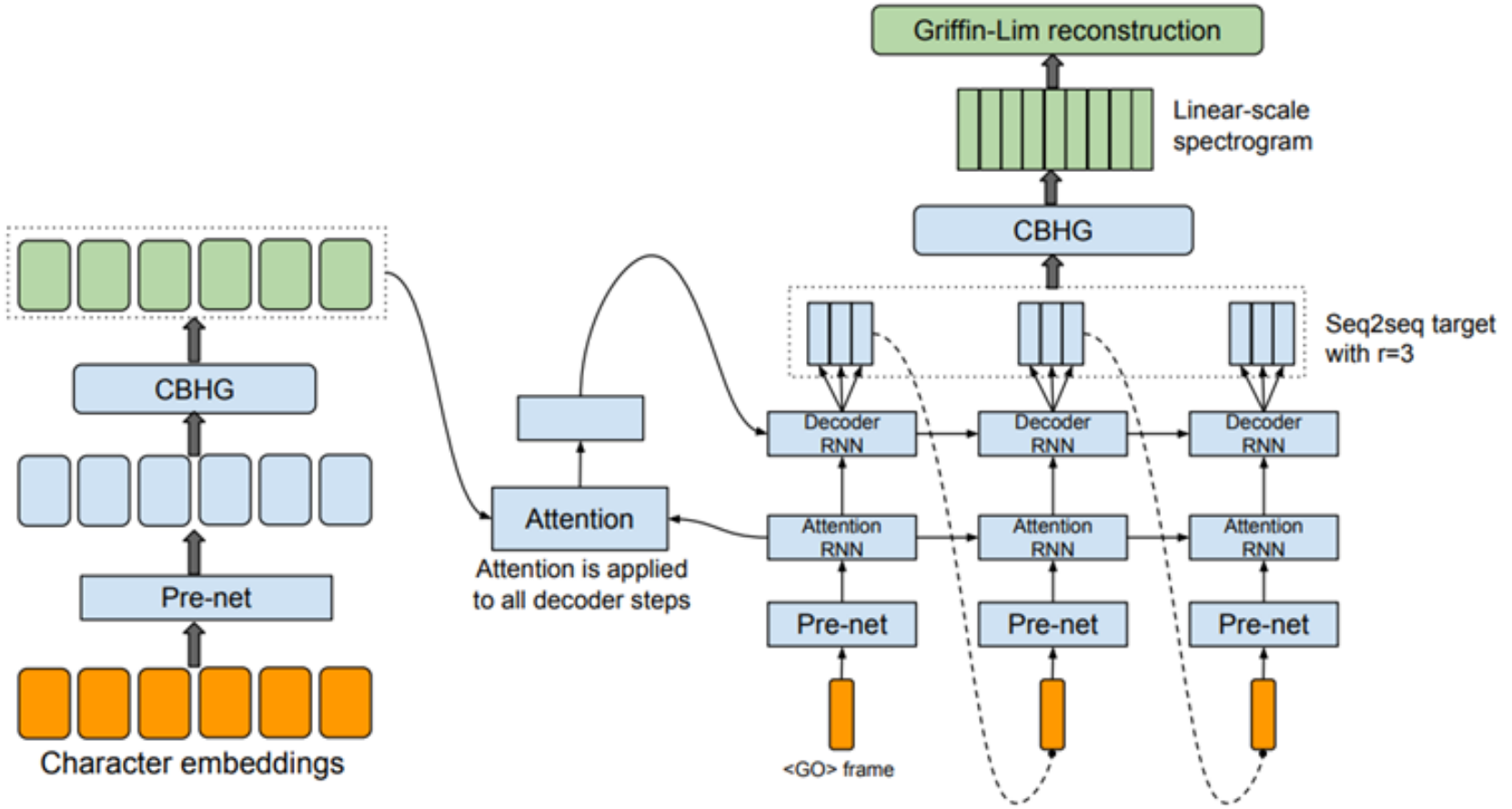
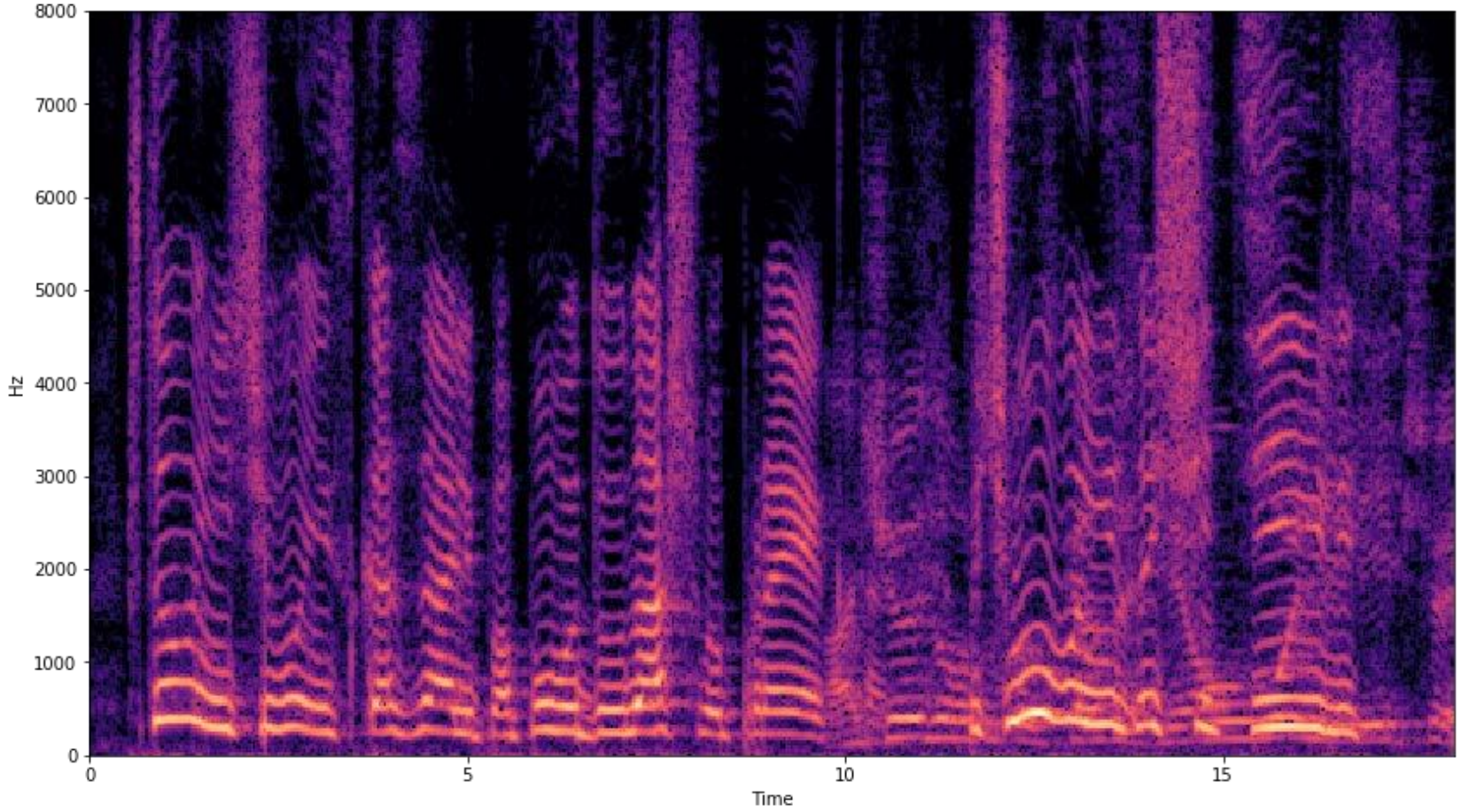


Figure 1: *Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech.*

Спектрограмма

Получается с помощью преобразования Фурье на коротких фрагментах звукового сигнала



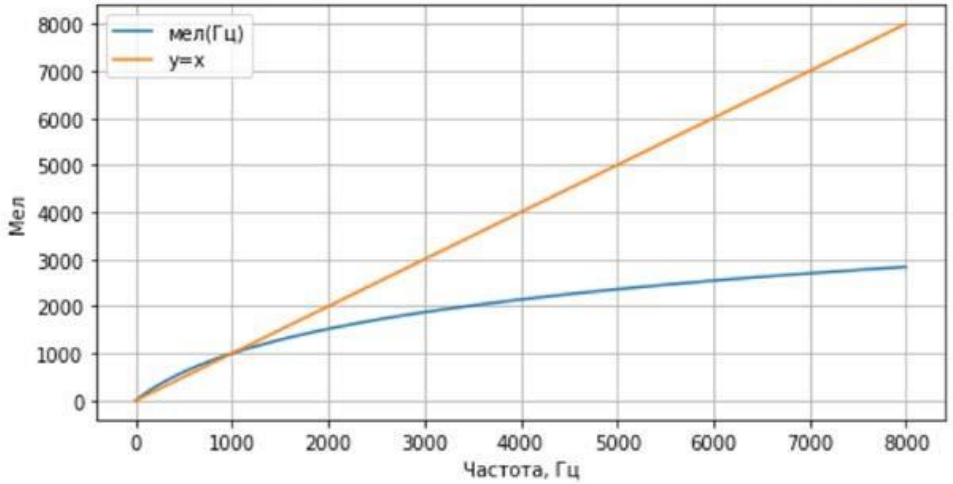
Мел-спектрограмма

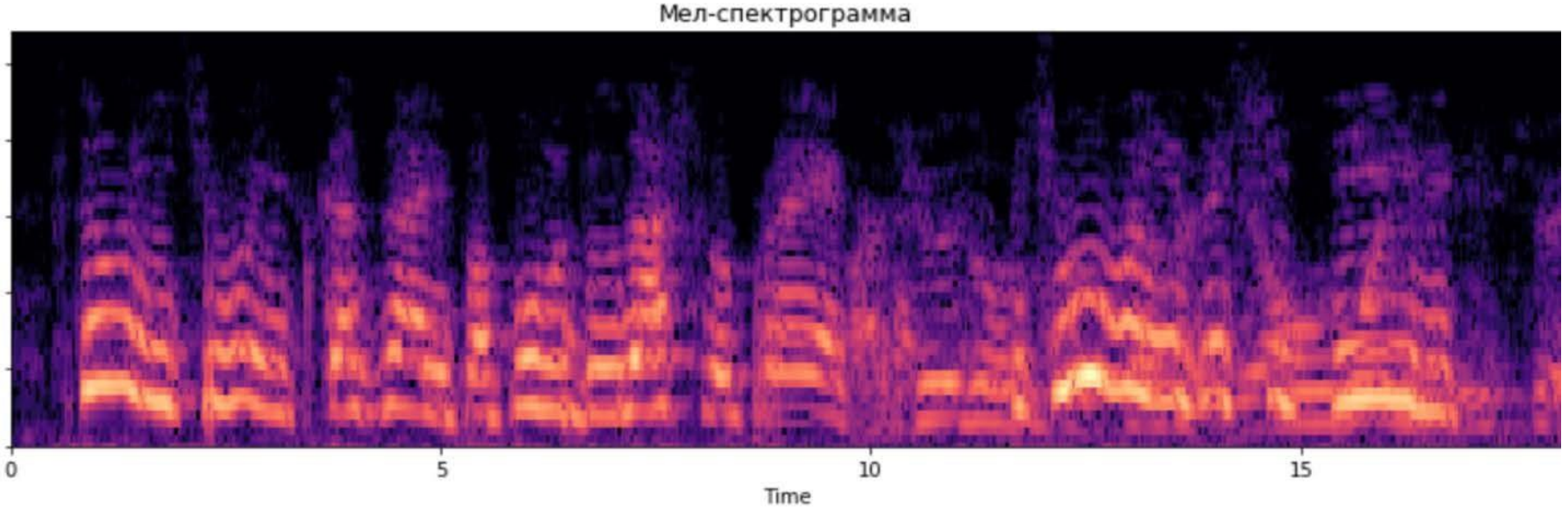
Мел – единица измерения, основана на психо-физиологическом восприятии звука человеком и логарифмически зависит от частоты.

Человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких.

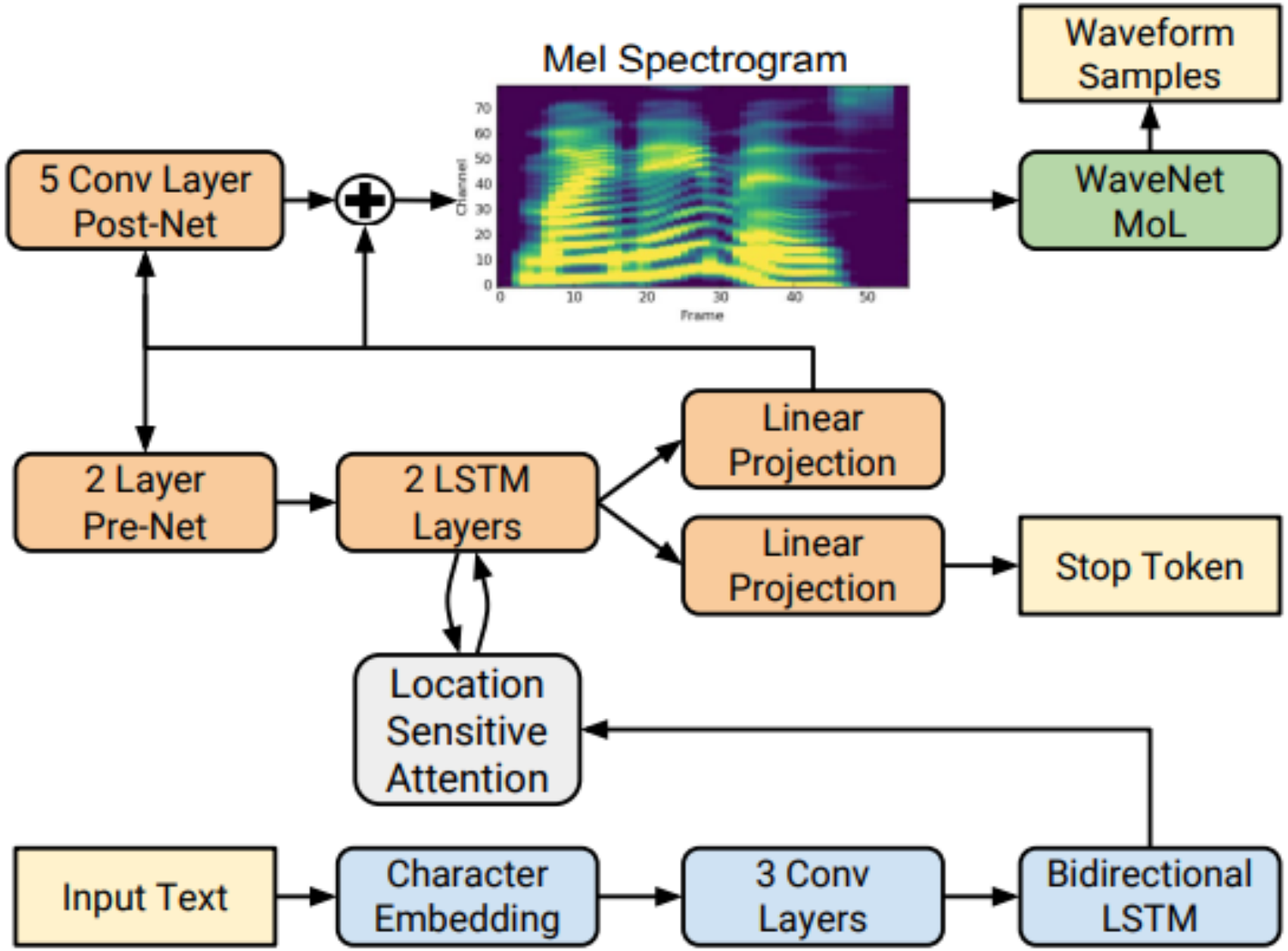
Мел-спектрограмма получается из спектрограммы с помощью формулы:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right)$$





Tacotron 2

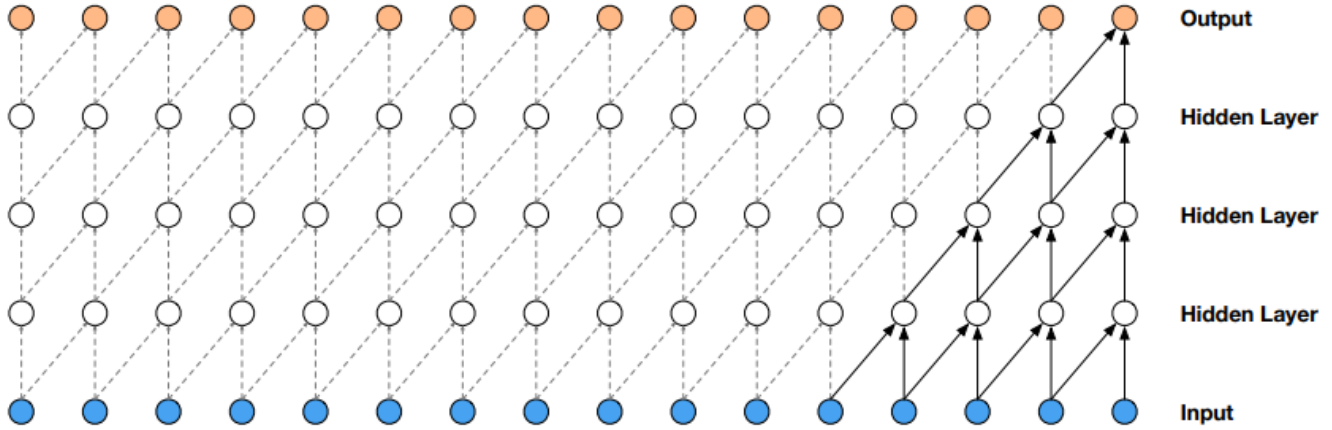

Mel Spectrogram
Waveform Samples
5 Conv Layer Post-Net
WaveNet MoL
Linear Projection
2 Layer Pre-Net
2 LSTM Layers
Linear Projection
Stop Token
Location Sensitive Attention
Input Text
Character Embedding
3 Conv Layers
Bidirectional LSTM

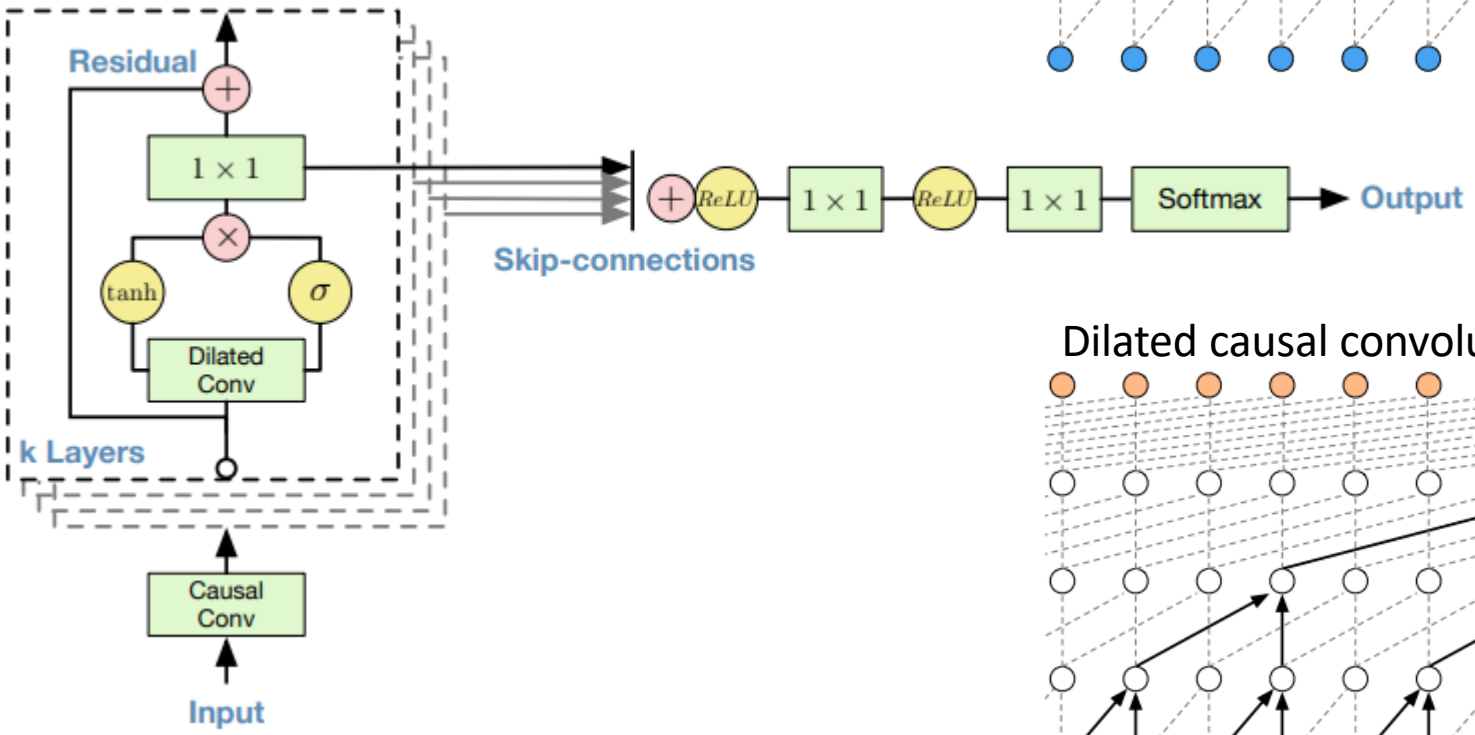
WaveNet

The joint probability of a waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$

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

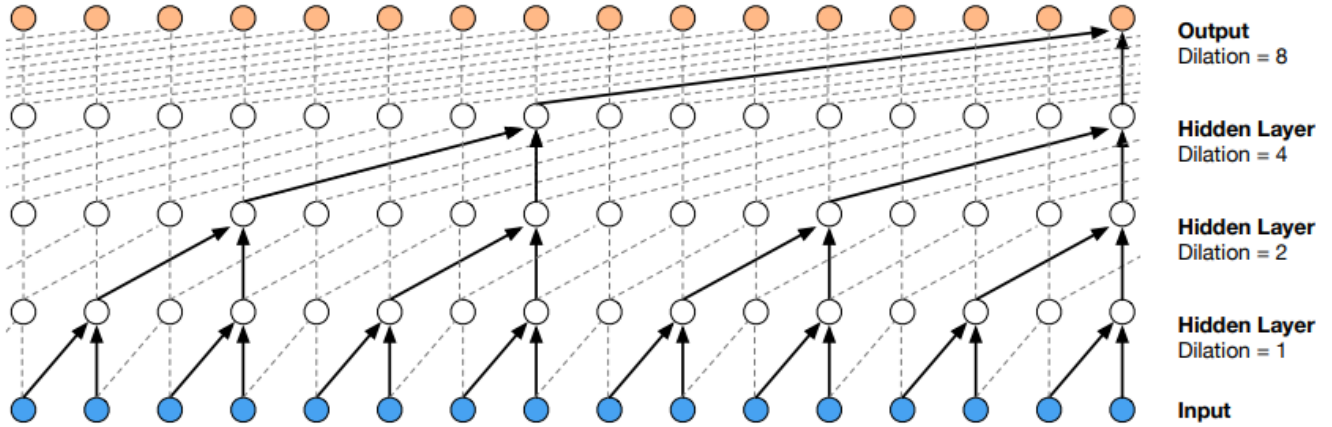
Causal convolutions





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

Dilated causal convolutions

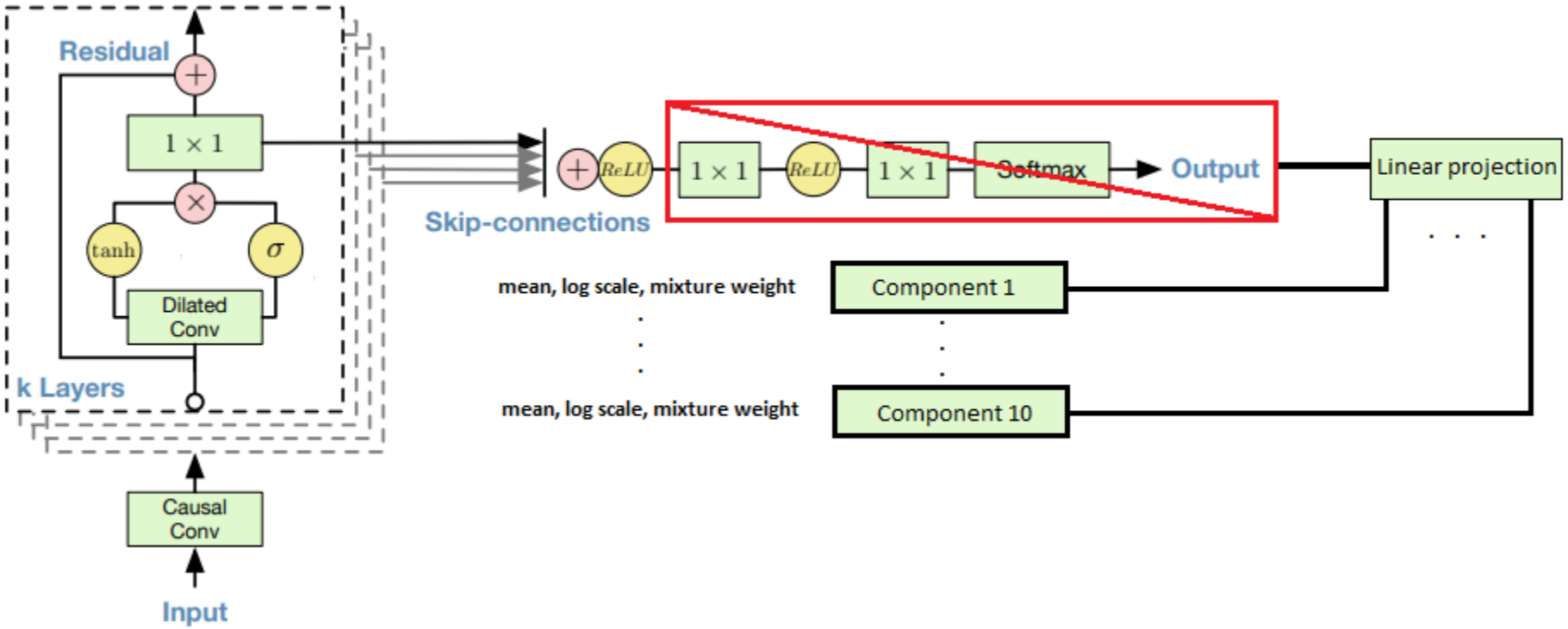


WaveNet: MoL

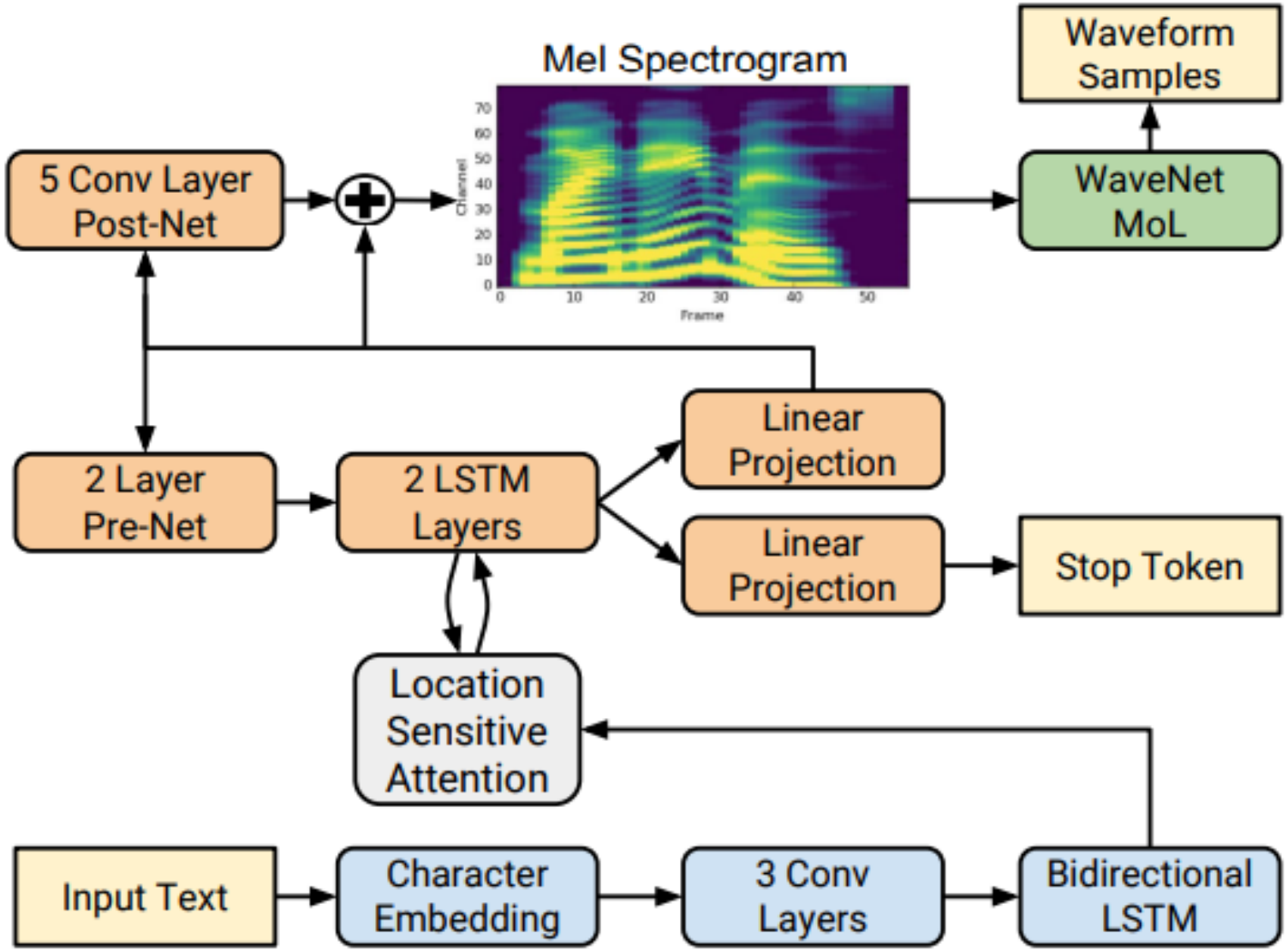
Теперь работаем с мел-спектрограммой вместо лингвистических признаков

Discretized 10-component mixture of logistic distributions (MoL) вместо Softmax

Функция потерь = NLLLoss от настоящего сэмпла

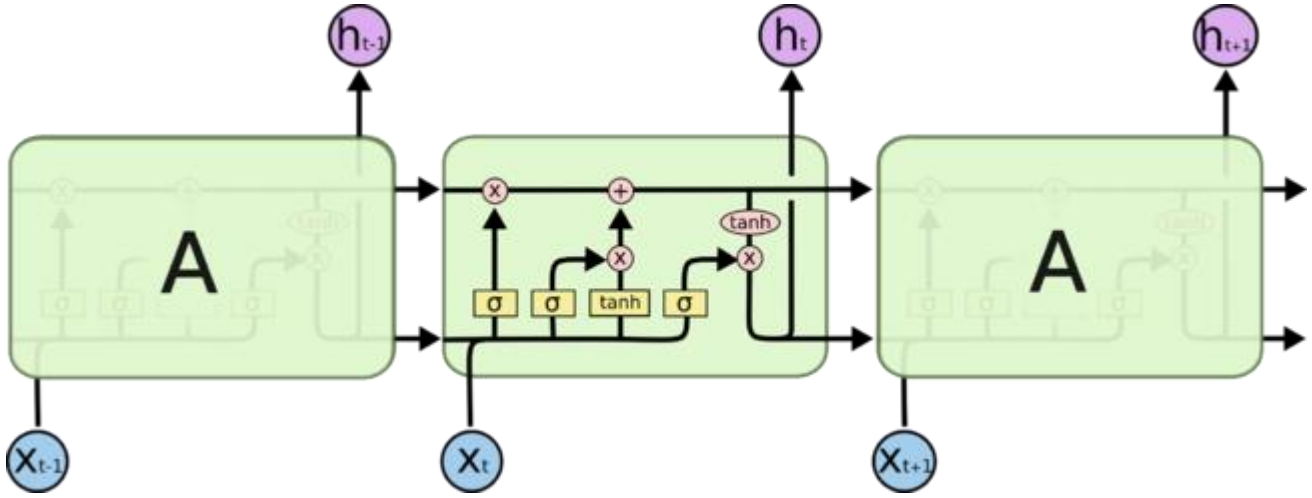


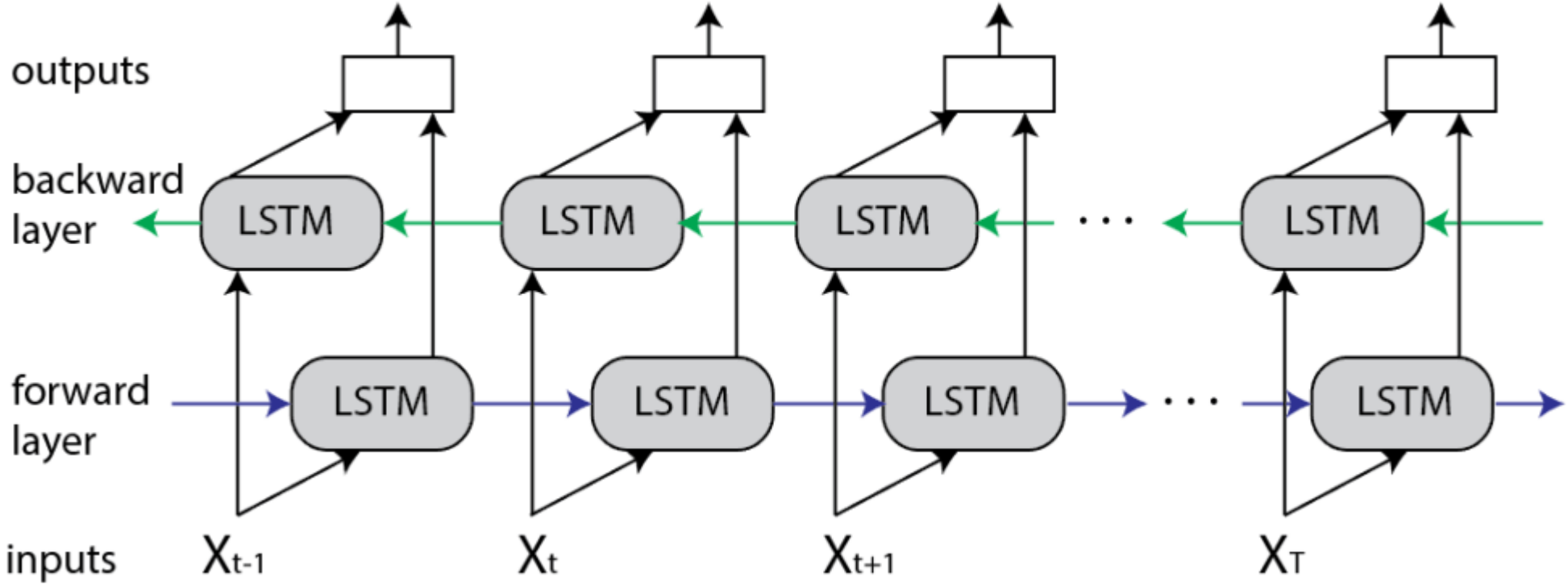
Tacotron 2


Mel Spectrogram
Waveform Samples
5 Conv Layer Post-Net
WaveNet MoL
Linear Projection
2 Layer Pre-Net
2 LSTM Layers
Linear Projection
Stop Token
Location Sensitive Attention
Input Text
Character Embedding
3 Conv Layers
Bidirectional LSTM

Bidirectional LSTM

- Проходим LSTM сначала в одну сторону потом обратно





Location Sensitive Attention

$$\alpha_i = Attend(s_{i-1}, \alpha_{i-1}, h): \quad \alpha_{i,j} = \exp(e_{i,j}) \Big/ \sum_{j=1}^{L} \exp(e_{i,j})$$

$$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j$$

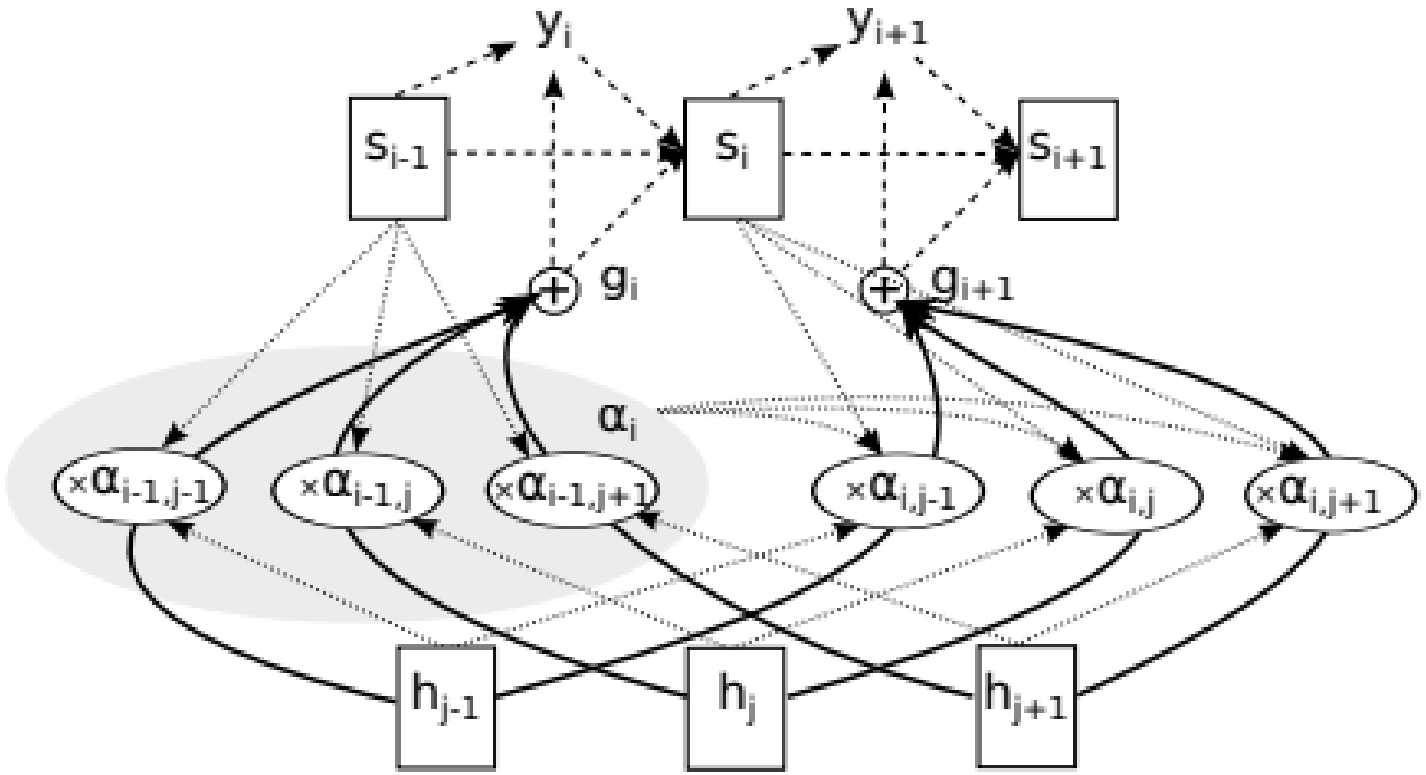
$$y_i \sim Generate(s_{i-1}, g_i),$$

$$s_i = Recurrency(s_{i-1}, g_i, y_i)$$

$$e_{i,j} = Score(s_{i-1}, h_j),$$

$$f_i = F * \alpha_{i-1}.$$

$$e_{i,j} = w^\top \tanh(W s_{i-1} + V h_j + U f_{i,j} + b)$$



Эксперименты и результаты:

Обучение

- Для обеих нейросетей используем *teacher-forcing,* оптимизируем Adam
- Для FPN используем batch size = 64 и один GPU
- Для WaveNet используем batch size = 128 и 32 GPU с синхронным обновлением
- Учимся на Internal US English dataset – 24.6 часа речи от одного спикера
- Текст нормализован

Эксперименты и результаты:

Оценка

- MOS сравним с ground truth
- −0.270 ± 0.155 в side-by-side сравнении с человеческой речью по шкале от -3 до 3.

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)	$\mathbf{4.526 \pm 0.066}$

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

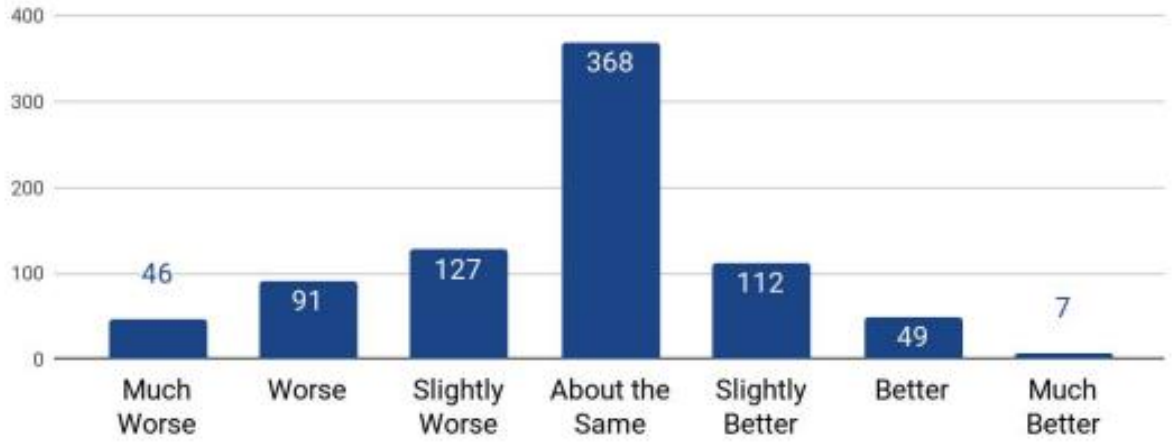


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

Эксперименты и результаты:

Ablation Studies

- Обучение и синтез WaveNet на ground truth и predicted сэмплах
- Разные промежуточные представления и вокодеры
- Разный размер WaveNet

Training	Synthesis: Predicted	Synthesis: Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

Table 2. Comparison of evaluated MOS for our system when WaveNet trained on predicted/ground truth mel spectrograms are made to synthesize from predicted/ground truth mel spectrograms.

System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	$\mathbf{4.526 \pm 0.066}$

Table 3. Comparison of evaluated MOS for Griffin-Lim vs. WaveNet as a vocoder, and using 1,025-dimensional linear spectrograms vs. 80-dimensional mel spectrograms as conditioning inputs to WaveNet.

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

Table 4. WaveNet with various layer and receptive field sizes.

Заключение

- Tacotron 2 - нейронная TTS модель, совмещающая seq2seq нейронную сеть для генерации мел-спектрограммы и модифицированный WaveNet вокодер
- Совмещает в себе парадигмы рекуррентных нейронных сетей, энкодер-декодер и механизм внимания для предсказания акустических признаков
- Модель не опирается не требует комплексного feature engineering-а
- Просодия Tacotron и качество аудио WaveNet позволяют в совокупности достичь state-of-the-art результата синтеза, близкого к настоящей человеческой речи
- https://google.github.io/tacotron/publications/tacotron2/ - примеры работы Tacotron 2