Автоматический синтез речи. Нейросетевой синтез

П. А. Холявин

p.kholyavin@spbu.ru

22.04.2024

Структура

1. Модуль текстового анализа
2. Акустическая модель
3. Вокодер

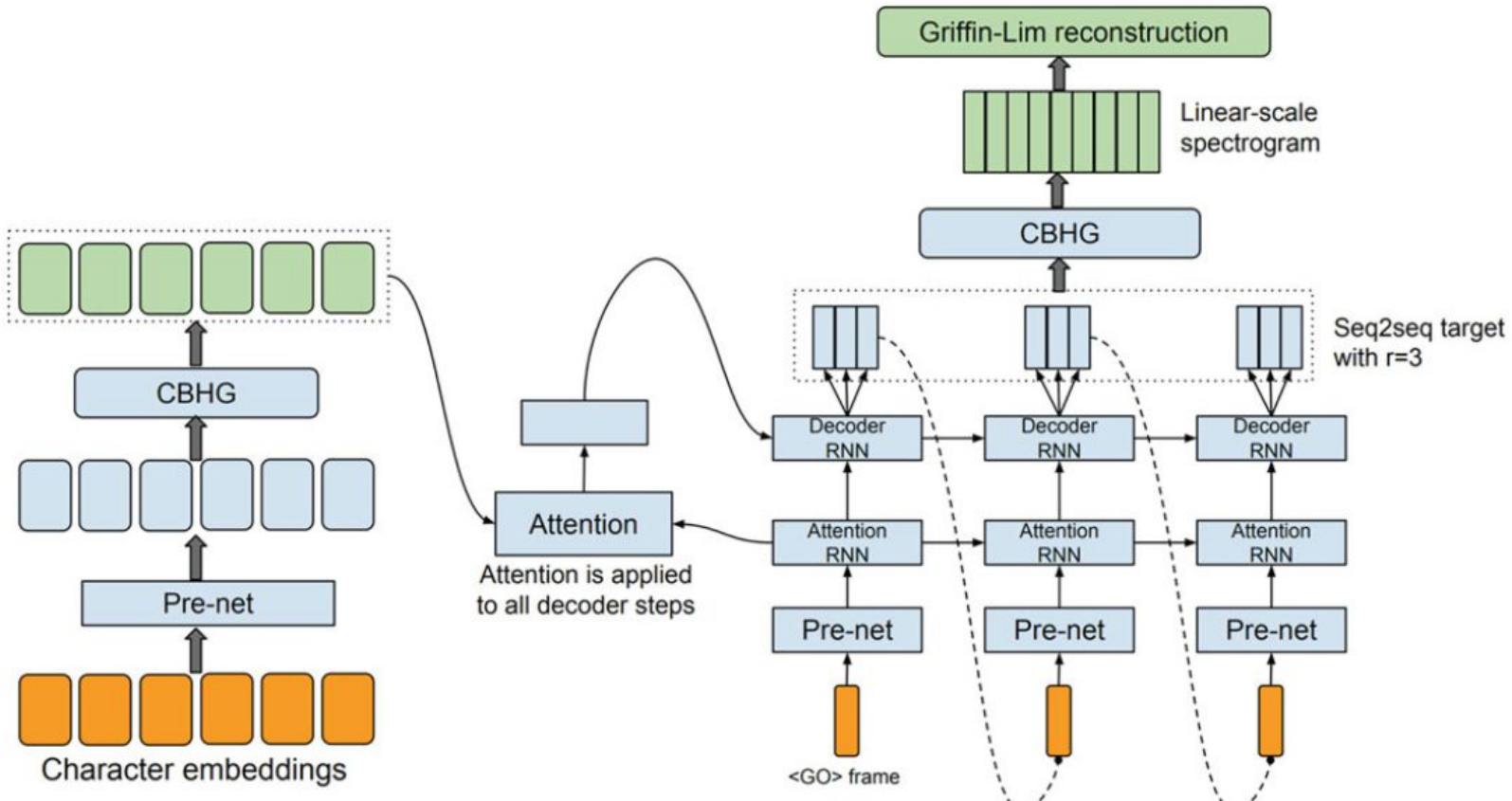
Акустические модели

На входе – последовательности букв или “фонем”

На выходе – акустические признаки высокой размерности (например, мел-спектрограммы)

Архитектура:
1. RNN
2. CNN
3. Трансформеры

RNN: Tacotron

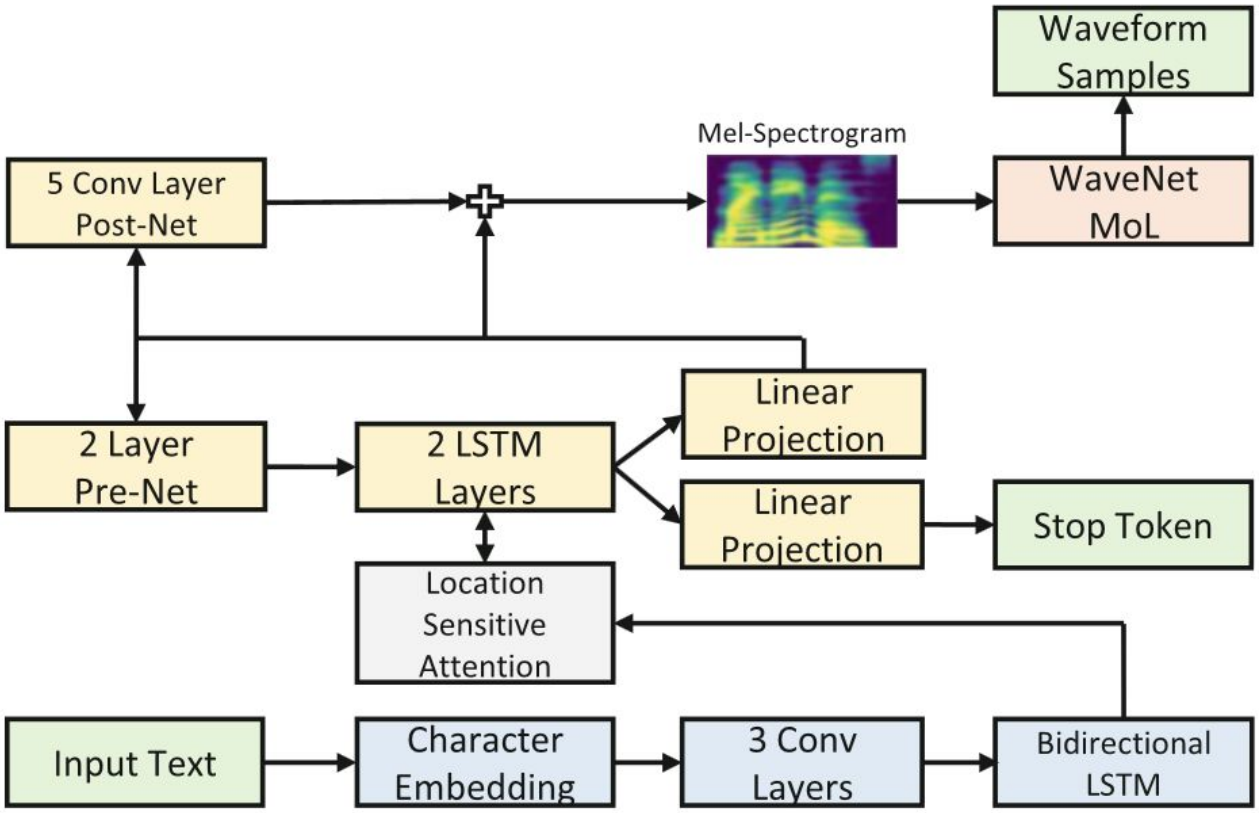


Алгоритм Гриффина-Лима

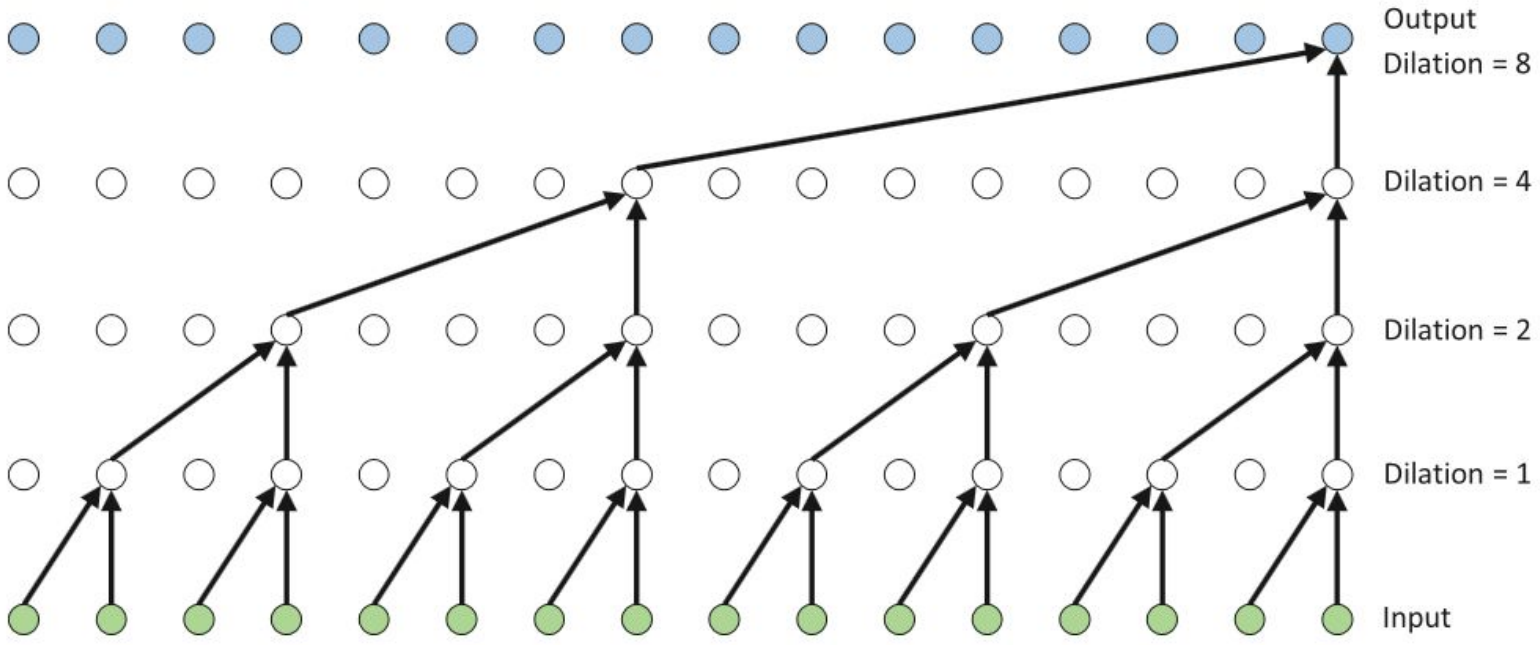
Algorithm 1 Griffin-Lim algorithm (GLA)

Fix the initial phase $\angle c_0$
Initialize $c_0 = s \cdot e^{i\angle c_0}$
Iterate for $n = 1, 2, \ldots$
$c_n = P_{C_1}\left(P_{C_2}(c_{n-1})\right)$
Until convergence
$x^* = \mathbf{G}^{\dagger} c_n$

LSTM: Tacotron 2

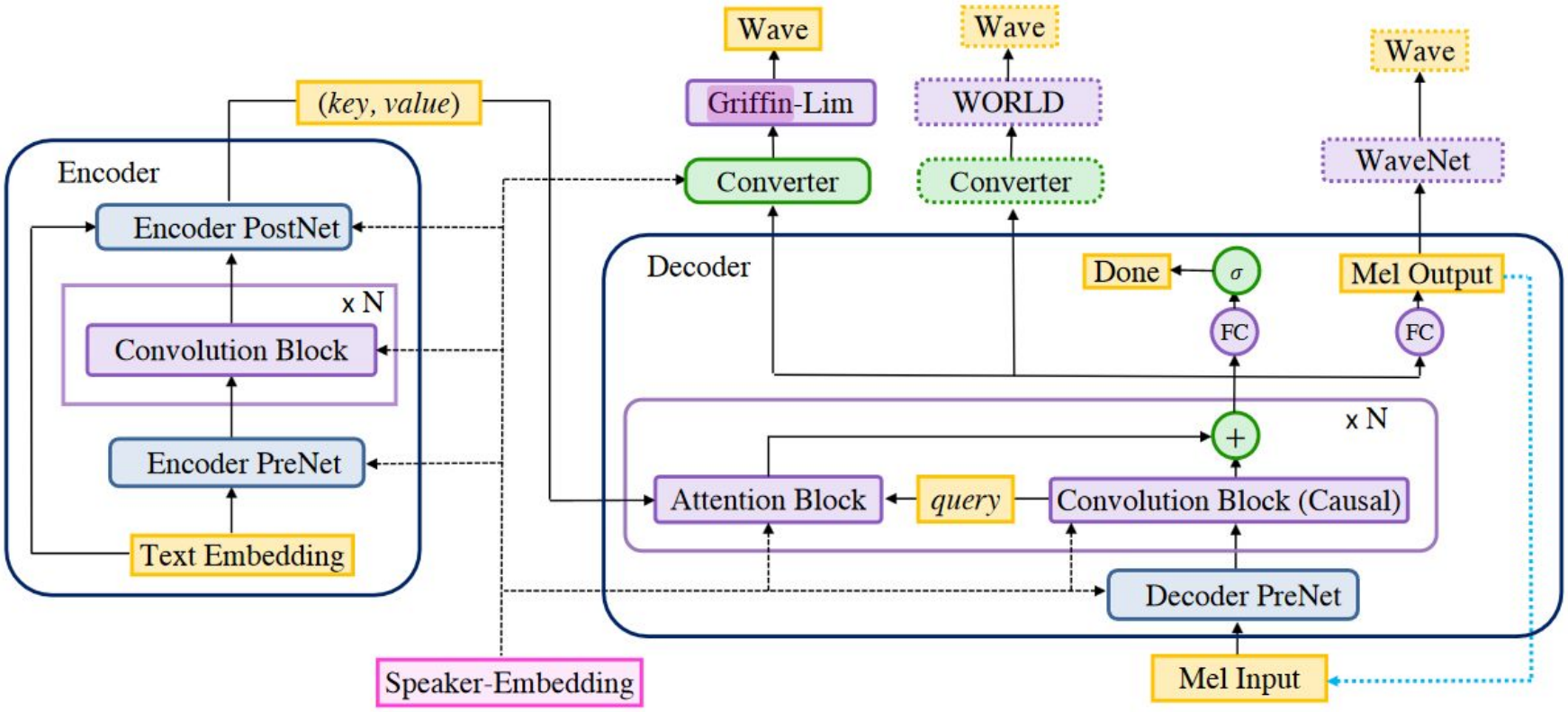


Вокодер WaveNet

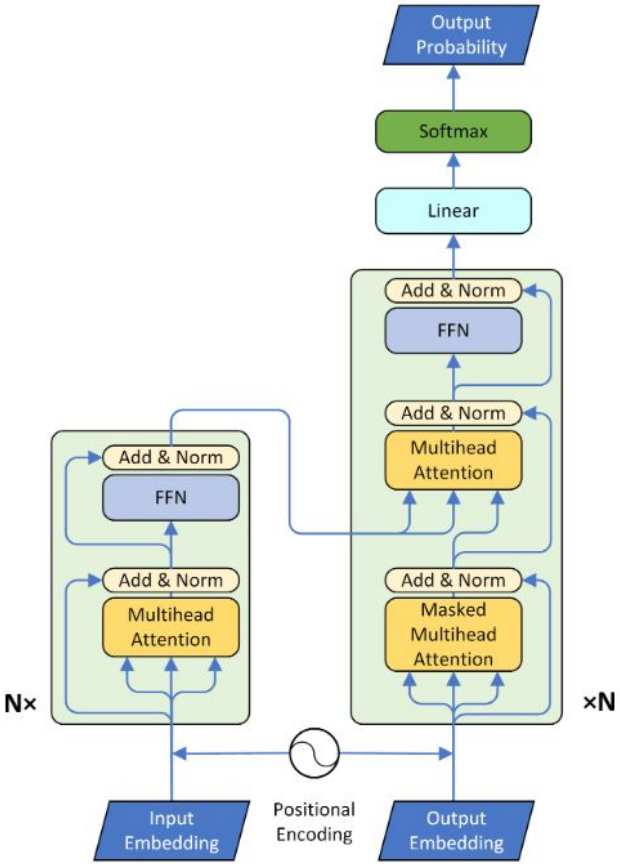


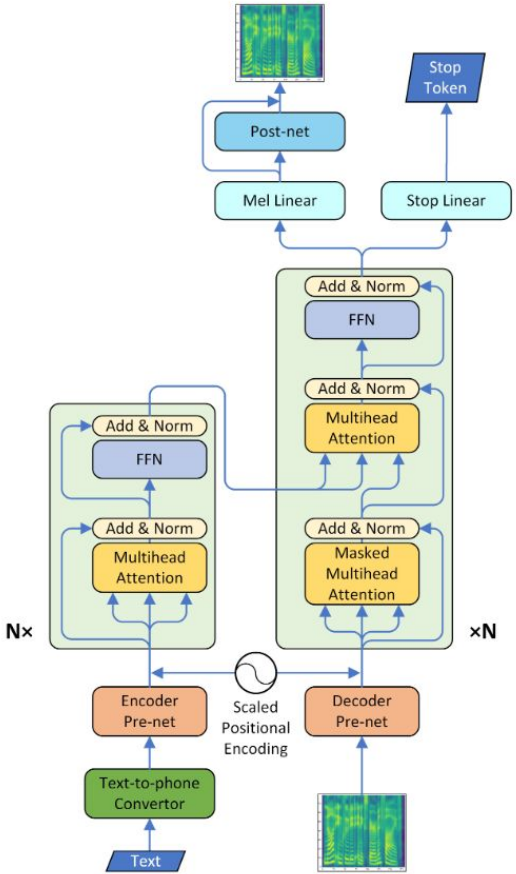
https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/

CNN: DeepVoice 3

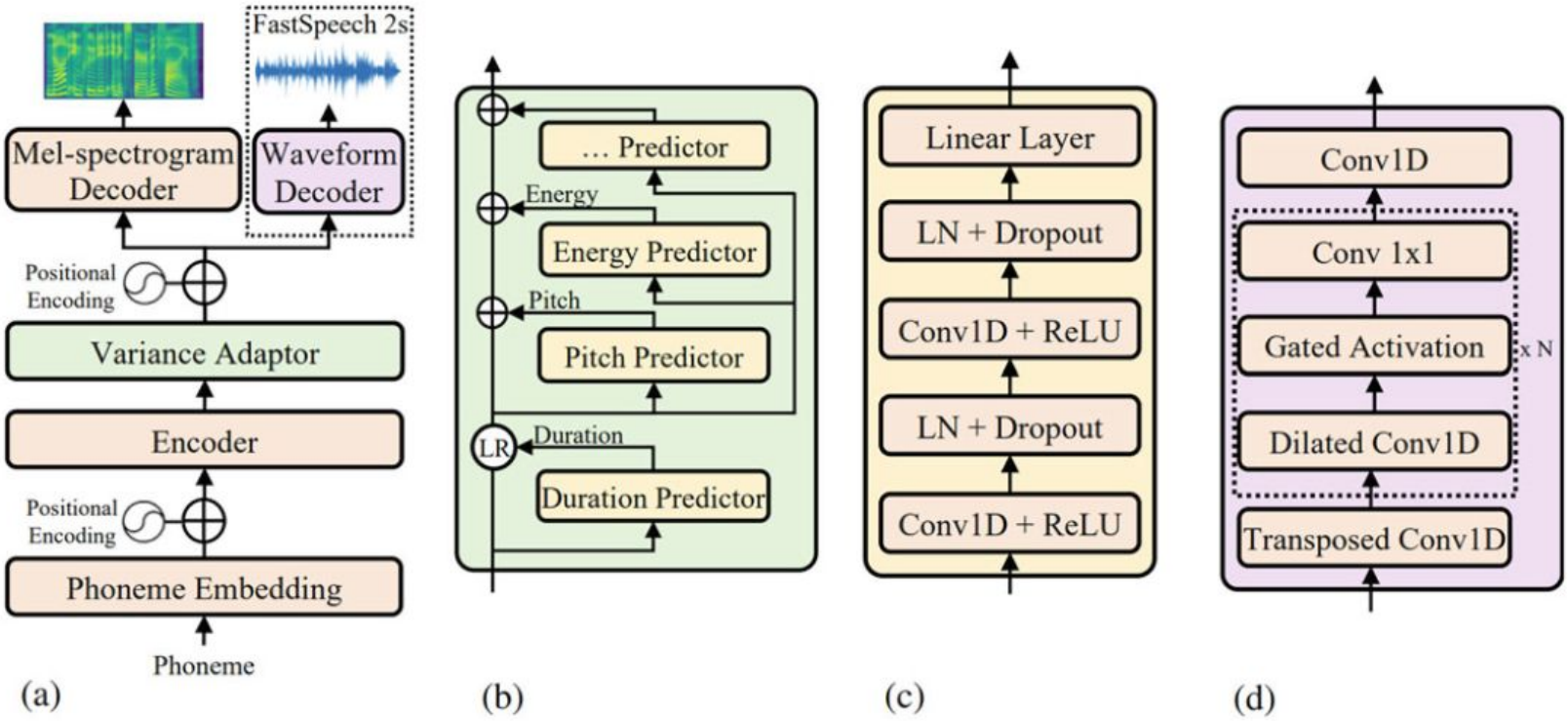


Transformer: FastSpeech





Transformer: FastSpeech 2



End-to-End модели

1) Совместное обучение акустической модели и вокодера

Char2Wav, Clarinet

2) Полностью параллельная структура

FastSpeech 2s, VITS

Спасибо за внимание!