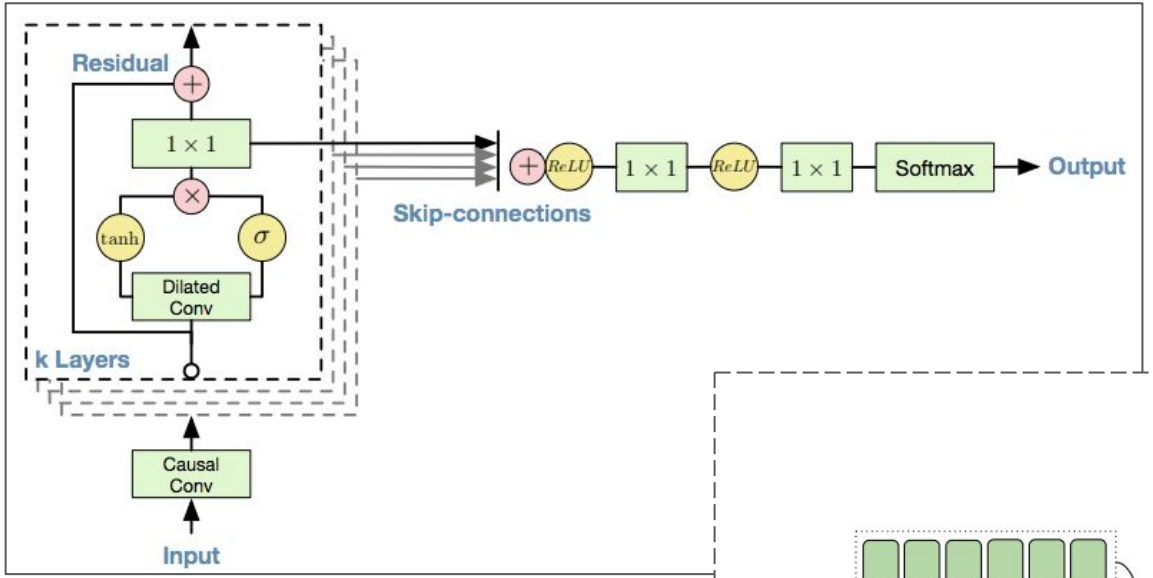
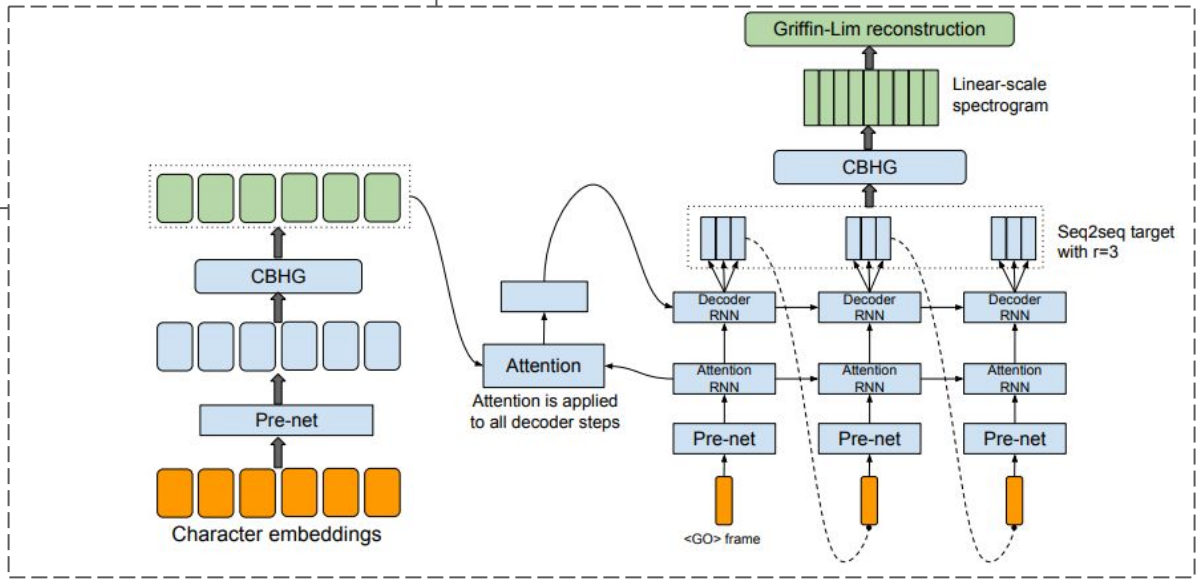
Tacotron 2

Предшествующие работы



- WaveNet

Tacotron -



Особенности этих подходов

- WaveNet:
  - WaveNet не может непосредственно выполнять работу с текстом и требует предварительную подготовку.
  + Качество результатов полученных от этой сети, схоже с человеческой речью

- Первый Tacotron:
  - Качество гораздо хуже, чем у WaveNet
  + Может полноценно работать с текстом

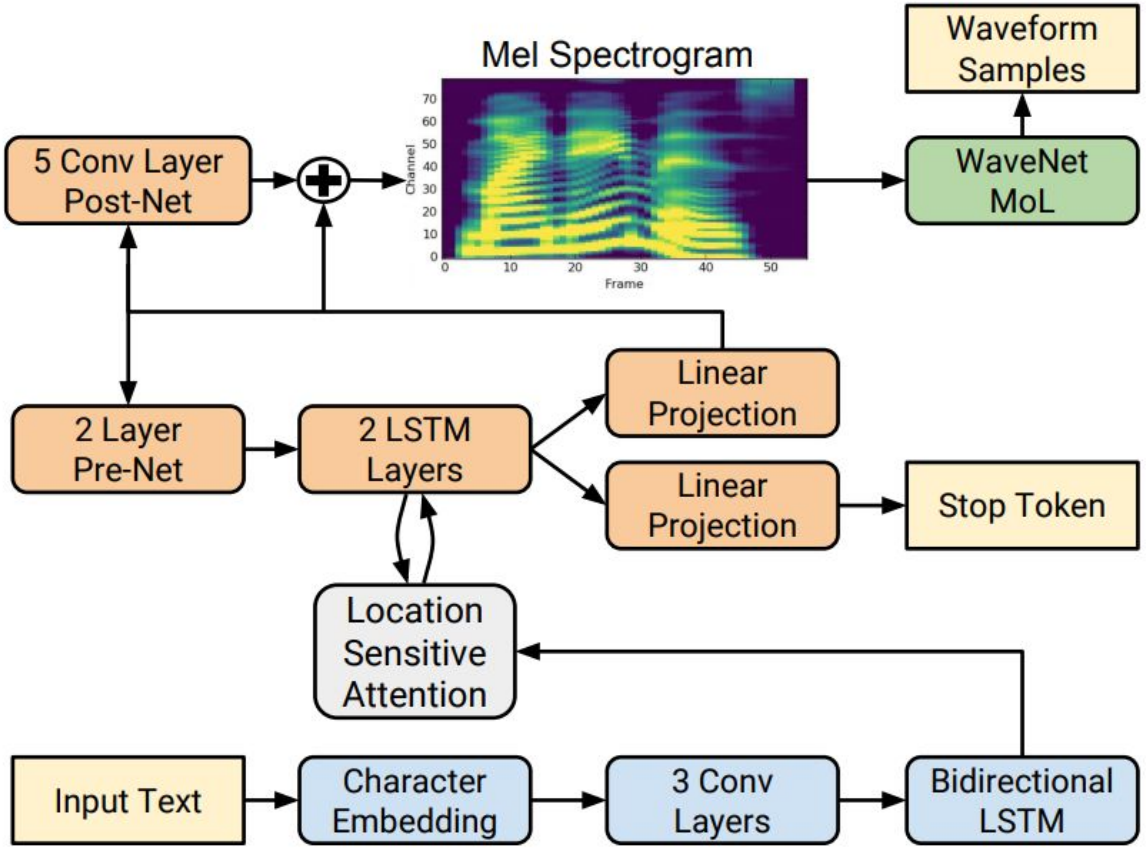
Tacotron 2

Основные части:

1) **Seq2Seq** - из текста в mel-spectrogram

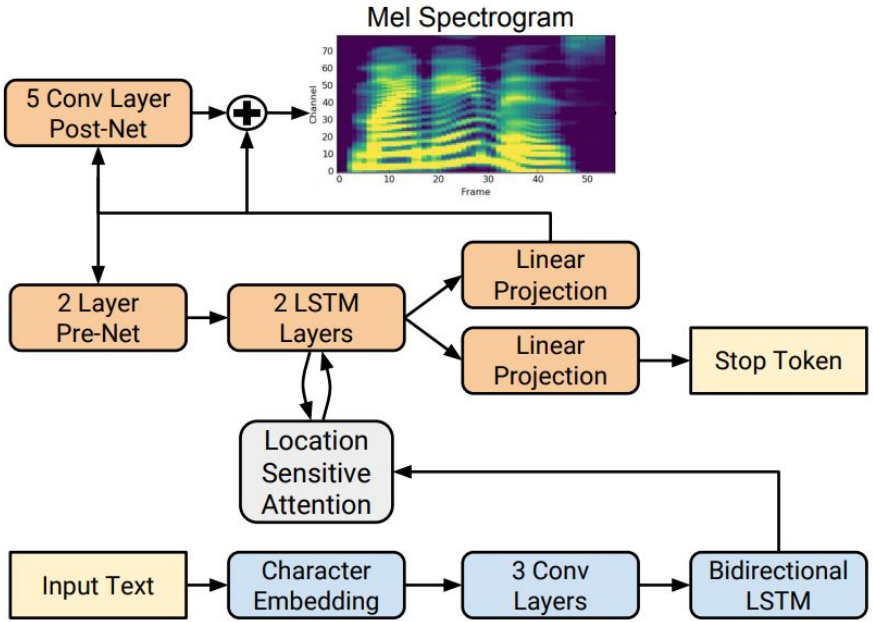
2) **WaveNet** модель для генерации самого звука

* Модели могут обучаться отдельно



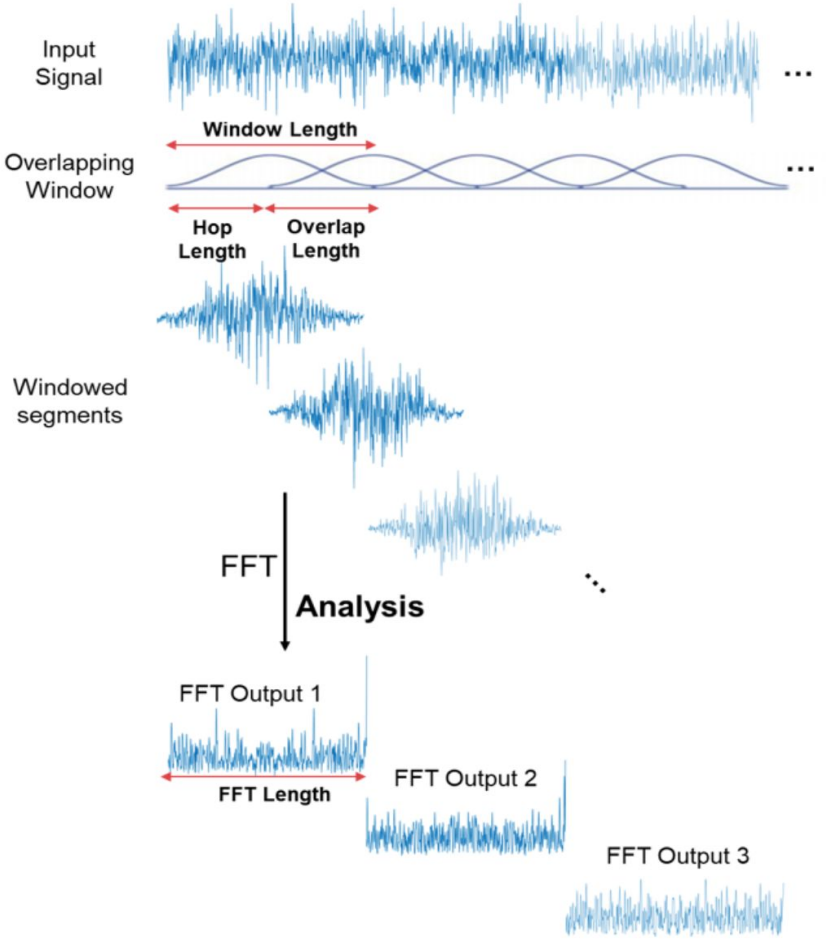
Часть Seq2Seq

- Структура проще, чем у первого такотрона.
- Из текста на входе генерируем мел спектрограмму
- Устройство этой части: encoder-attention-decoder



Mel Spectrogram

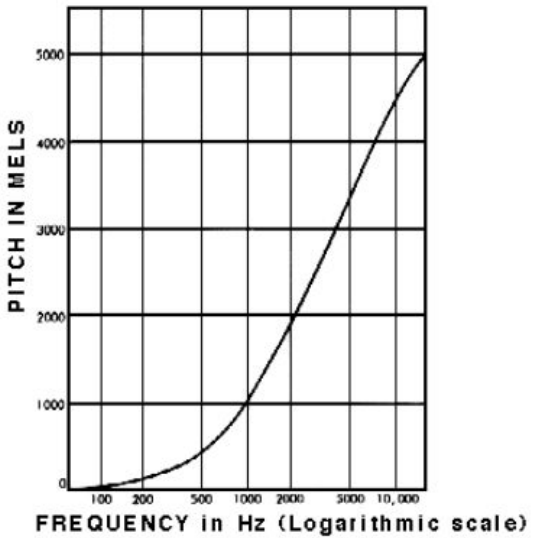
- Хотим перевести сигнал от декодера в функцию зависимости от частоты
- Используем Оконное преобразование Фурье
- Для этого берем окно Ханна
- window length - 50ms, hop length - 12.5 ms



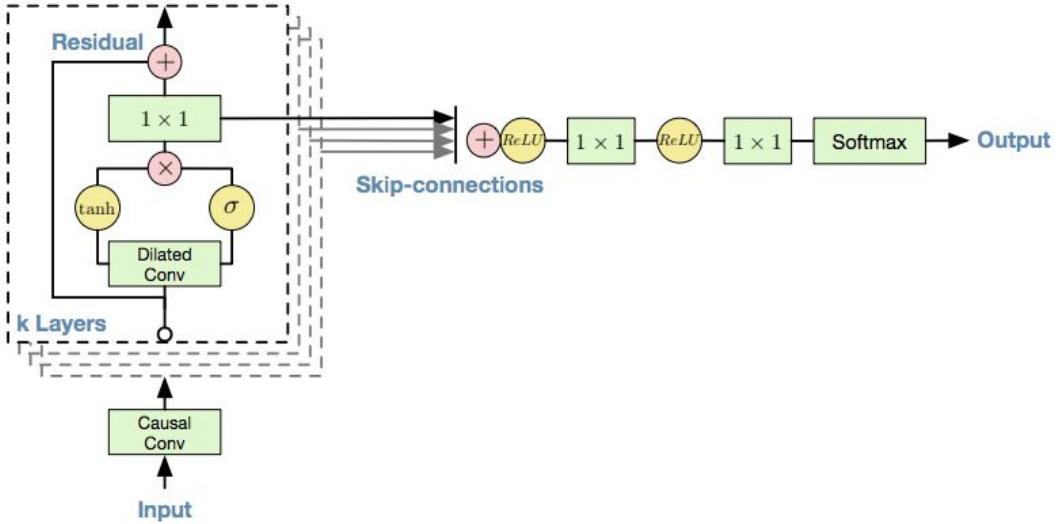
Mel Spectrogram

- Легче тренировать с MSE

- Человек лучше слышит разницу на низких частотах, нежели на высоких



WaveNet



Что изменили?

- Уменьшили в dilated conv число слоев на треть

- Заменили softmax на MoL (Mixture of logistics)

Результаты

- MOS (mean opinion score) - несколько (обычно 8) критиков ставят оценку тому, насколько речь похожа на человеческую по шкале от 1 до 5 с шагом 0.5.

- 100 примеров от каждого алгоритма

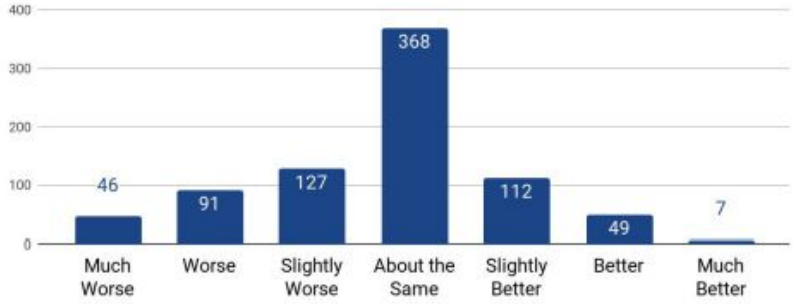
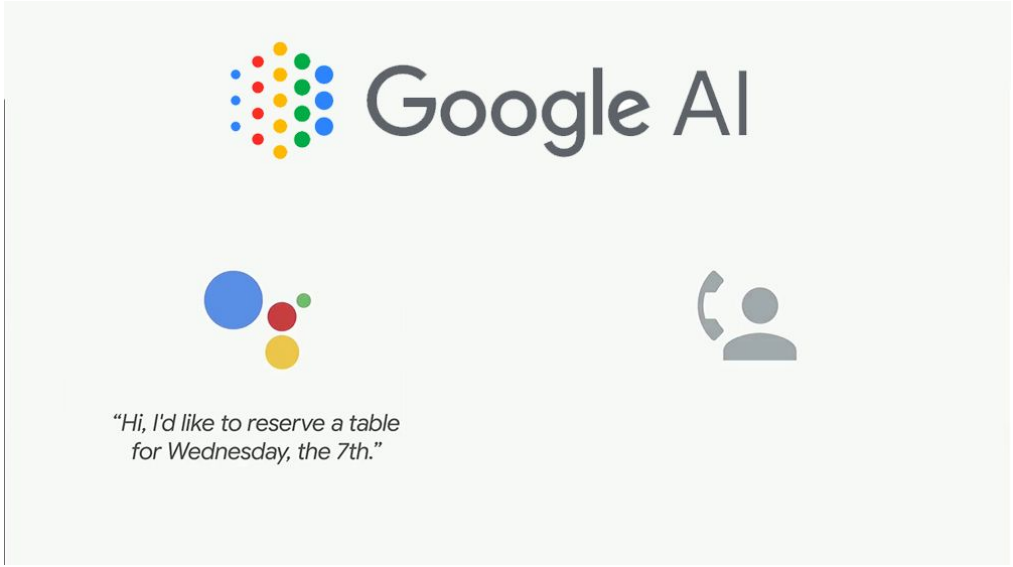


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

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)	$\mathbf{4.526 \pm 0.066}$

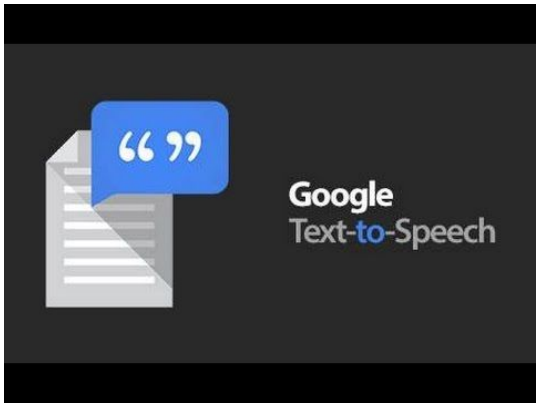
Применение

- В голосовых ассистентах



- Озвучка текста, в случаях когда сам человек не может

- Переводчики



Полезные ссылки

- https://github.com/NVIDIA/tacotron2/blob/master/model.py
- https://arxiv.org/pdf/1712.05884.pdf
- https://google.github.io/tacotron/publications/tacotron2/index.html
- https://google.github.io/tacotron/publications/tacotron2/index.html
- https://deepmind.com/blog/article/wavenet-generative-model-raw-audio