Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

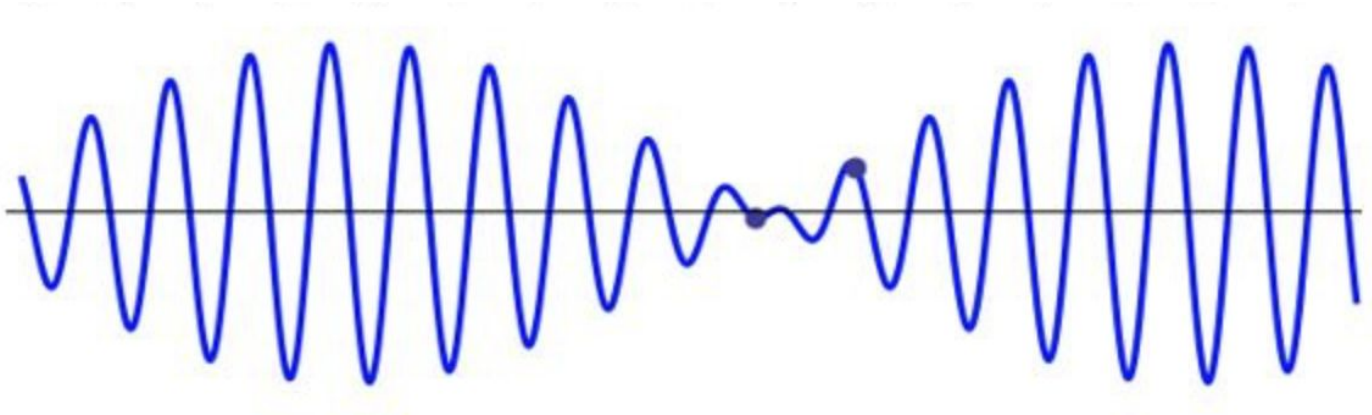
Сидоров Дмитрий

План

- Напоминание про хранение звуков в компьютере
- Напоминание про Text-To-Speech и WaveNet
- Описание метрики MOS (Mean Opinion Score)
- Tacotron
- Недостатки Tacotron
- Tacotron 2
- Результаты

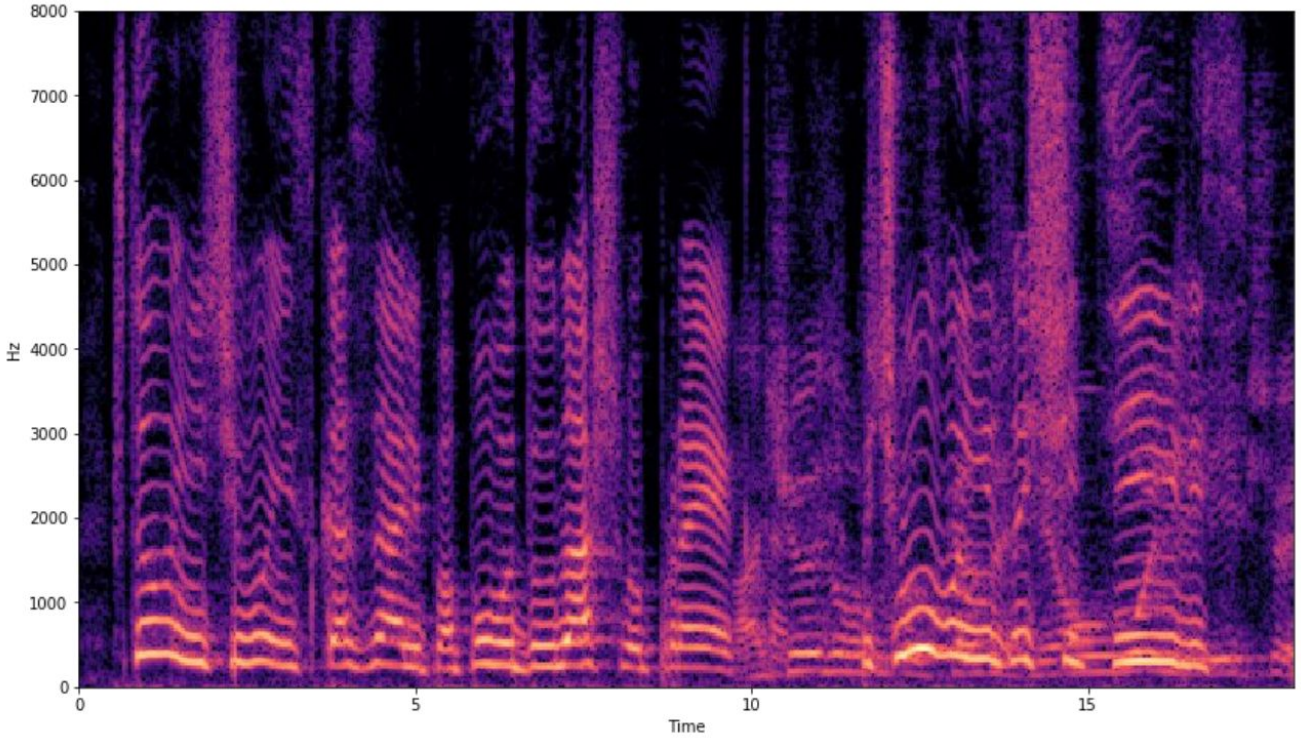
Хранение звуков в компьютере

- Звук – это волна => можно фиксировать амплитуды и хранить как последовательность чисел
- Обычно используются 16-битные числа



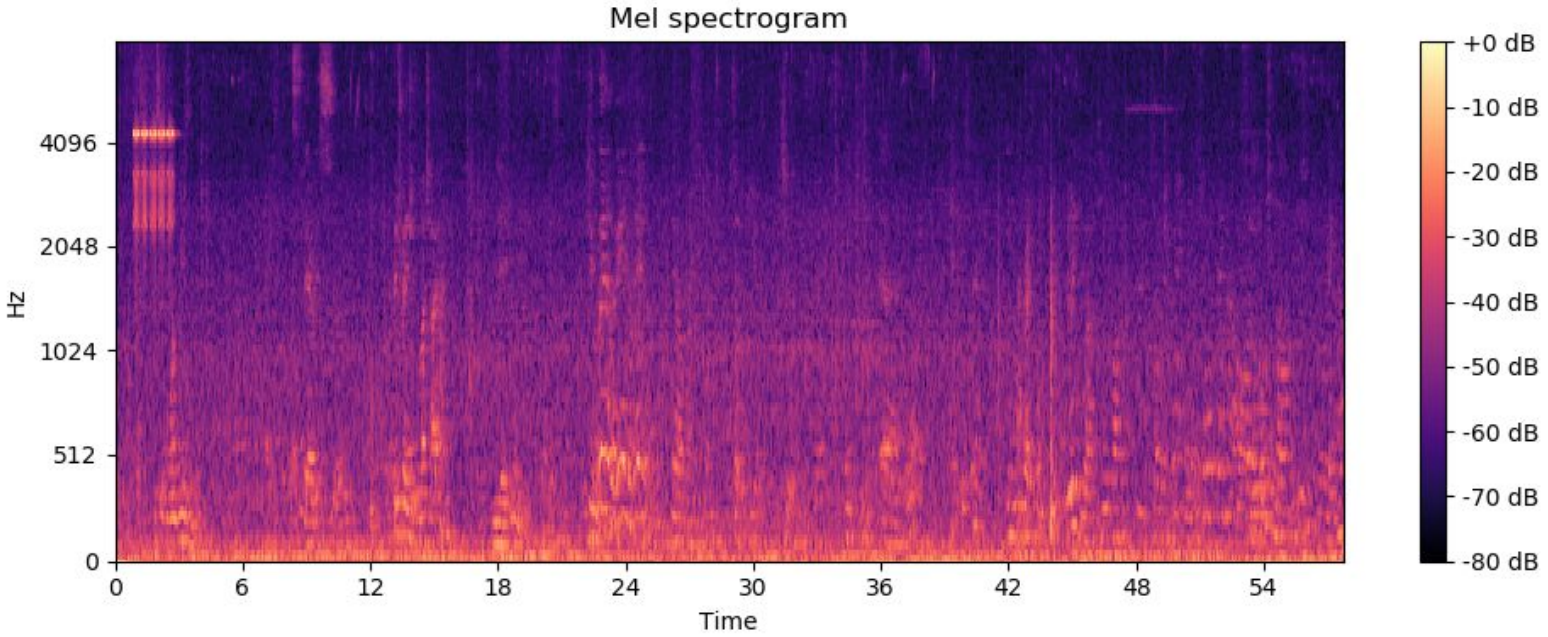
Спектрограммы

- Показывают зависимость амплитуды от времени и частоты



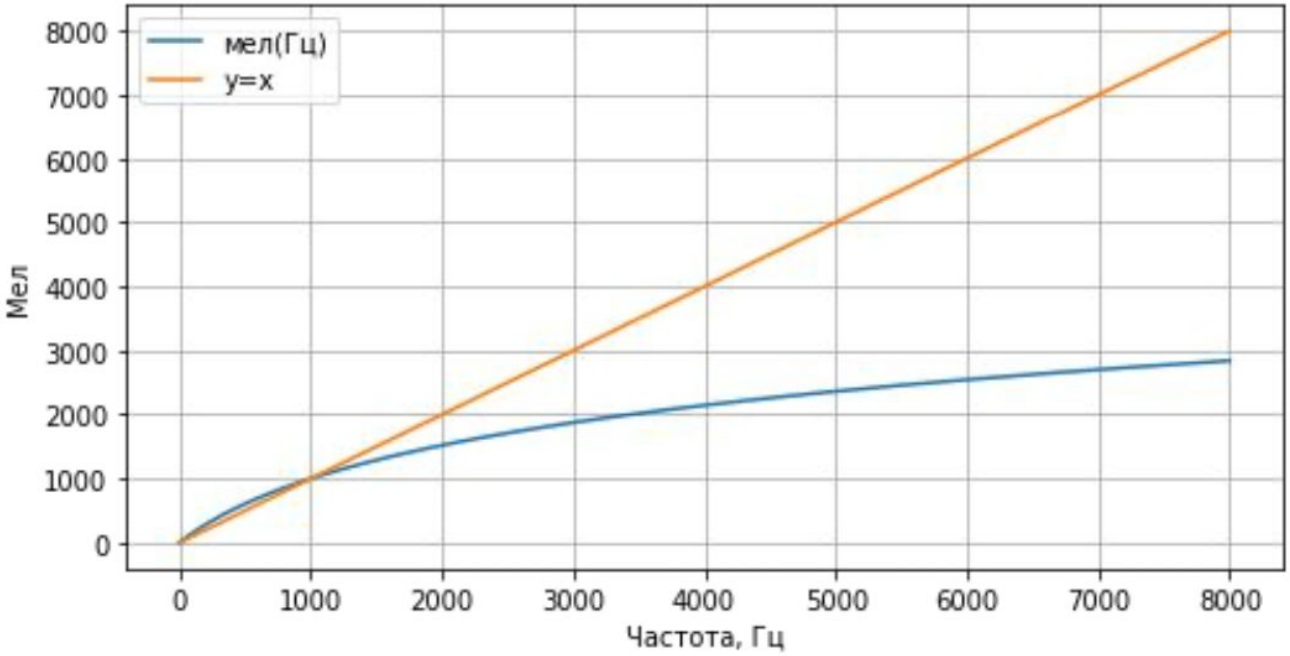
Мел-спектрограммы

- Человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких
- Мел - психофизическая единица высоты звука



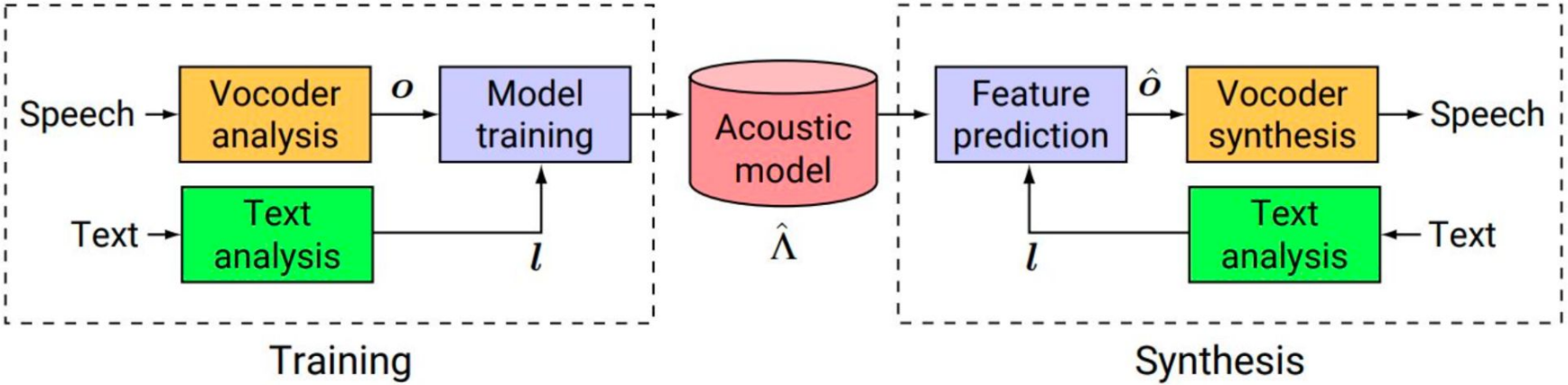
Мел-спектрограммы

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



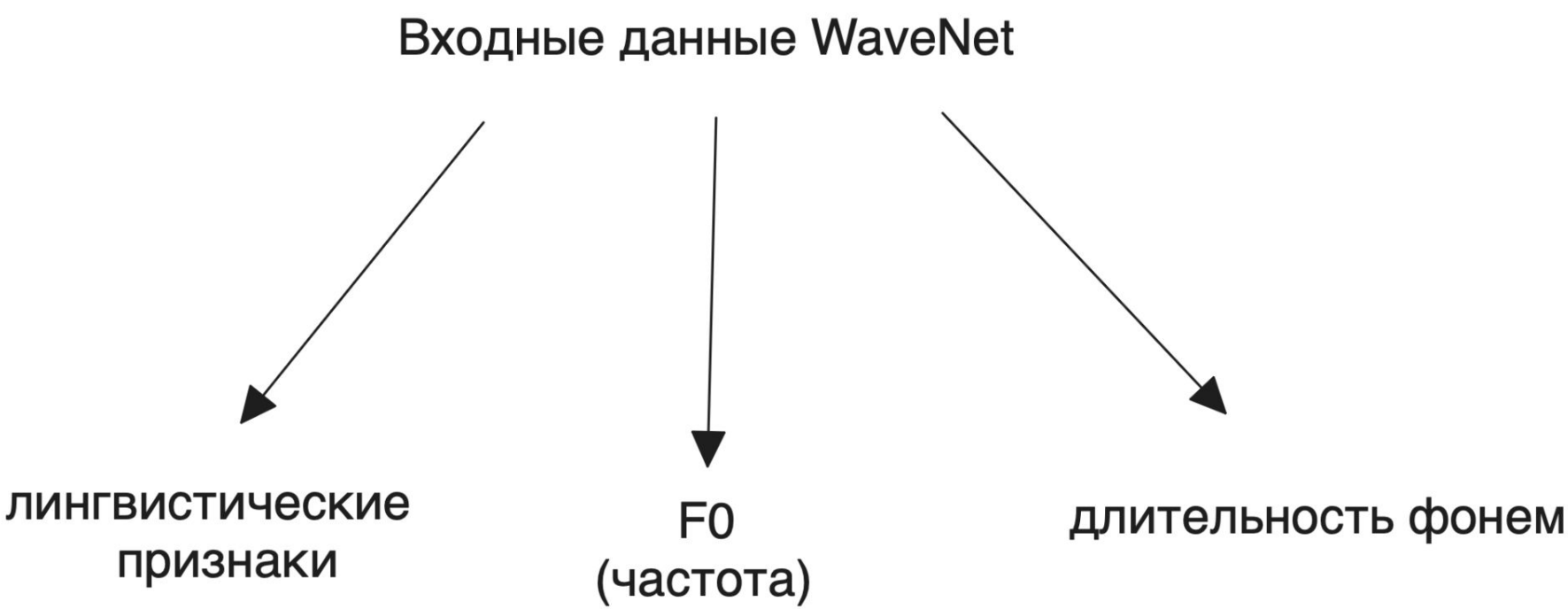
Text-To-Speech

- Задача – сгенерировать речь по тексту
- Есть две компоненты: анализ текста и синтез речи



WaveNet

- Предсказывает вероятности, что амплитуда в момент времени t примет каждое из возможных значений, если известны значения амплитуд в предыдущие моменты времени
- Для предсказаний используется SoftMax слой


Входные данные WaveNet
лингвистические признаки
F0 (частота)
длительность фонем

Из чего складывается оценка MOS

- Естественность речи
- Чистота и отсутствие артефактов
- Динамический диапазон
- и так далее

Оценки ставятся от 1 до 5 с шагом 0.5

MOS для WaveNet и Tacotron 2

- WaveNet – 4.34
- Tacotron 2 – 4.53
- Профессиональный диктор – 4.58

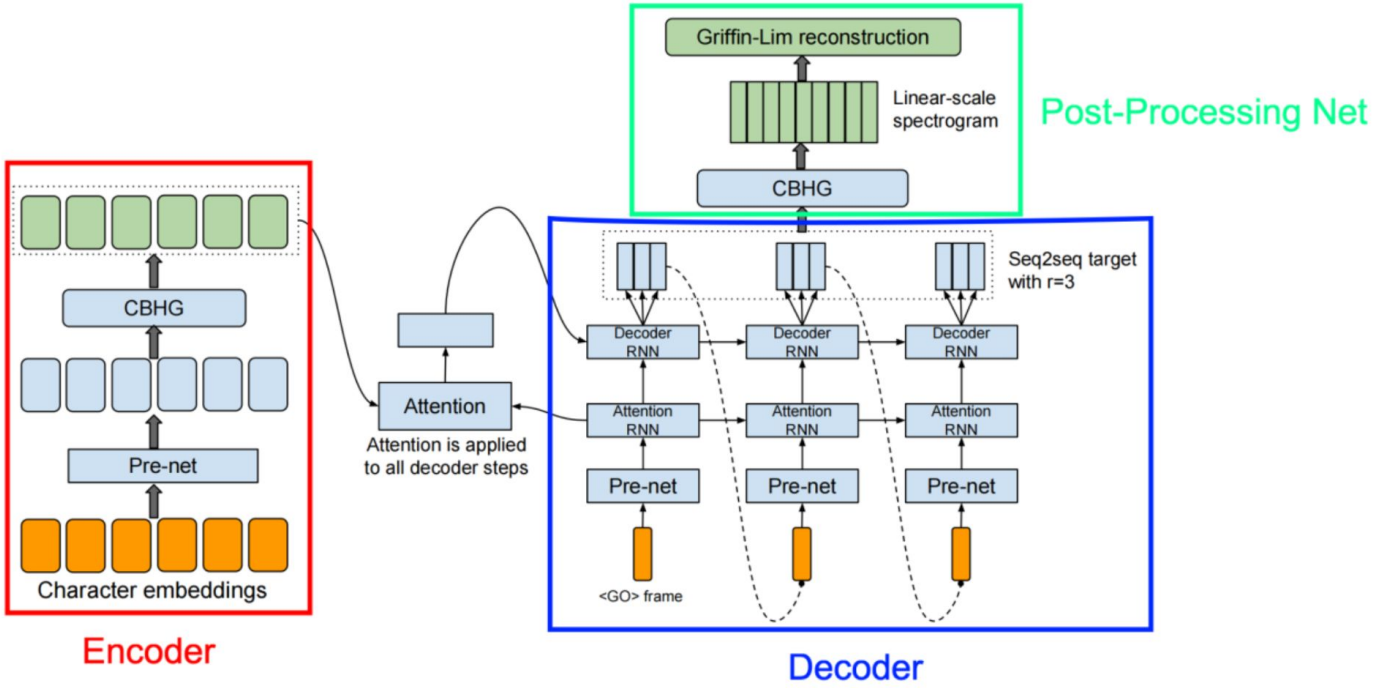
Сравнение Tacotron 2 с диктором можно посмотреть здесь:
https://disk.yandex.ru/d/Q_cQ0LcG1vS3ig

Tacotron (2017)

- Работает не с лингвистическими признаками, а с текстом
- Вход – текст
- Выход – mel-спектрограмма (далее обрабатывается голосовой моделью)
- MOS 4.0

Проблема – mel-спектрограммы не идеально озвучиваются

Схема работы Tacotron



Источник: https://anwarvic.github.io/speech-synthesis/Tacotron

Алгоритм Гриффина-Лима
(Алгоритм восстановления сигнала по его фазовой информации)

```python
def griffin_lim(spectrogram, n_iter=hp.n_iter):
    # Создаем копию спектрограммы для последующих манипуляций
    x_best = copy.deepcopy(spectrogram)

    # Итеративный процесс восстановления сигнала
    for i in range(n_iter):
        # Применяем обратное преобразование Фурье (ISTFT) к текущей лучшей спектрограмме
        x_t = librosa.istft(x_best, hp.hop_length, win_length=hp.win_length, window="hann")

        # Вычисляем новую спектрограмму с использованием текущего приближения сигнала
        est = librosa.stft(x_t, hp.n_fft, hp.hop_length, win_length=hp.win_length)

        # Вычисляем фазу для обновления лучшей спектрограммы
        phase = est / np.maximum(1e-8, np.abs(est))

        # Обновляем лучшую спектрограмму с использованием фазы
        x_best = spectrogram * phase

    # Применяем обратное преобразование Фурье (ISTFT) к финальной лучшей спектрограмме
    x_t = librosa.istft(x_best, hp.hop_length, win_length=hp.win_length, window="hann")

    # Извлекаем вещественную часть, чтобы получить реальный сигнал
    y = np.real(x_t)

    return y
```

Спектрограммы (выводы)

- Не содержат информацию о фазе (содержат только об амплитуде)
- Чтобы генерировать аудио, нужно знать фазу
- Фаза зависит от времени
- Фазу можно восстановить с помощью алгоритма Гриффина-Лима
- Но эта оценка не позволяет достичь идеального качества

Tacotron 2

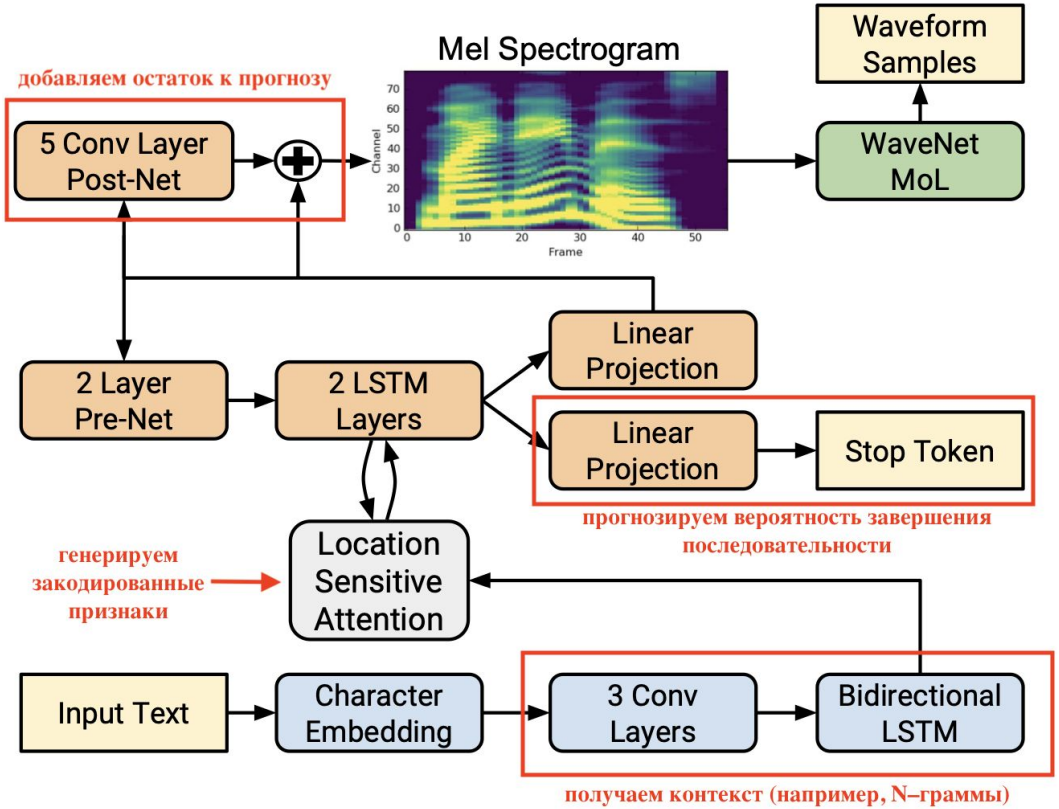


Fig. 1. Block diagram of the Tacotron 2 system architecture.

Location Sensitive Attention

- принимает на вход закодированную последовательность от LSTM (Long short-term memory)
- берёт произвольную последовательность и возвращает вектор контекста фиксированной длины
- уменьшает вероятность пропустить подпоследовательность
- нужен, чтобы декодер последовательно обрабатывал данные

подробнее:
https://aicurious.io/glossary/location%20sensitive%20attention

Декодер Tacotron 2

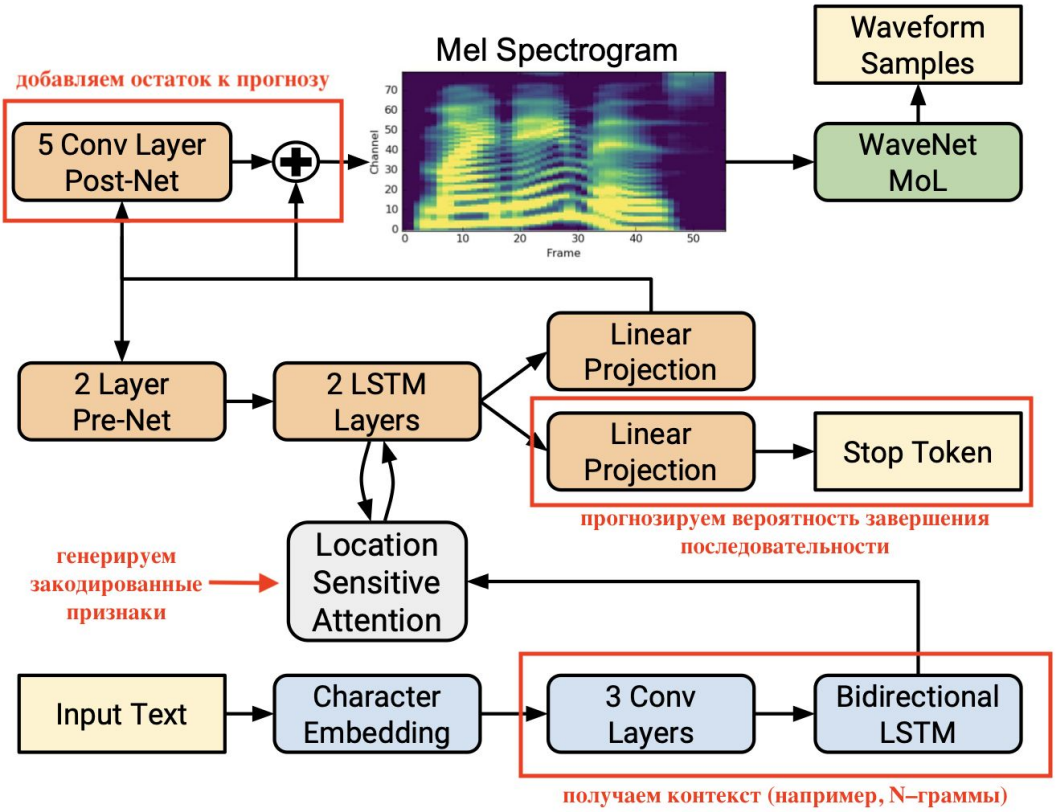


Fig. 1. Block diagram of the Tacotron 2 system architecture.

- Каждое следующее предсказание зависит от предыдущего
- На каждом шаге берём предыдущее предсказание, пропускаем через Pre-Net. Её выход объединяется с LSA и передаётся в 2 LSTM Layers, далее получаем набор вероятностей для спектограмм
- Stop Token используется, чтобы динамически определять, когда прекратить генерацию
- Post-Net пытается предсказать остаток, который мы упустили на предыдущих шагах

Модификации WaveNet

- Вместо предсказания дискретных сегментов предсказываем распределение вероятностей
- Используем 10-компонентную смесь логистических распределение (MoL)
- По факту из задачи классификации пришли к задаче регрессии
- WaveNet теперь генерирует аудио по mel-спектограмме

Результаты

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.

Сравнение с обучением на прогнозируемых мел-спектограммах

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.

Сравнение с линейными спектограммами

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.

Источники

- https://arxiv.org/pdf/1712.05884.pdf
- https://anwarvic.github.io/speech-synthesis/Tacotron
- https://habr.com/ru/companies/speechpro/articles/358816/
- https://aicurious.io/glossary/location%20sensitive%20attention