Московский государственный университет
Факультет вычислительной математики и кибернетики

Реализация алгоритма преобразования текста в речь в реальном времени с применением вейвлет преобразования

Студент группы М210:

Киреев Никита Сергеевич

Научный руководитель:

д. т. н., профессор Сухомлин Владимир Александрович

Научный консультант:

Ильюшин Евгений Альбинович

Москва
2020

Постановка задачи

1. Выполнить аналитический обзор литературы по теме трансляции текста в речь.
2. Рассмотреть современные архитектуры моделей синтеза речи и провести их сравнительный анализ.
3. Исследовать возможность улучшения качества синтезированной речи при помощи вейвлет преобразований.
4. Разработать алгоритм синтеза речи с применением вейвлет преобразования.
5. Реализовать предложенный алгоритм и провести качественное сравнение с существующими решениями.

Синтез речи

Синтез речи - восстановление формы речевого сигнала по его параметрам

Виды синтеза речи

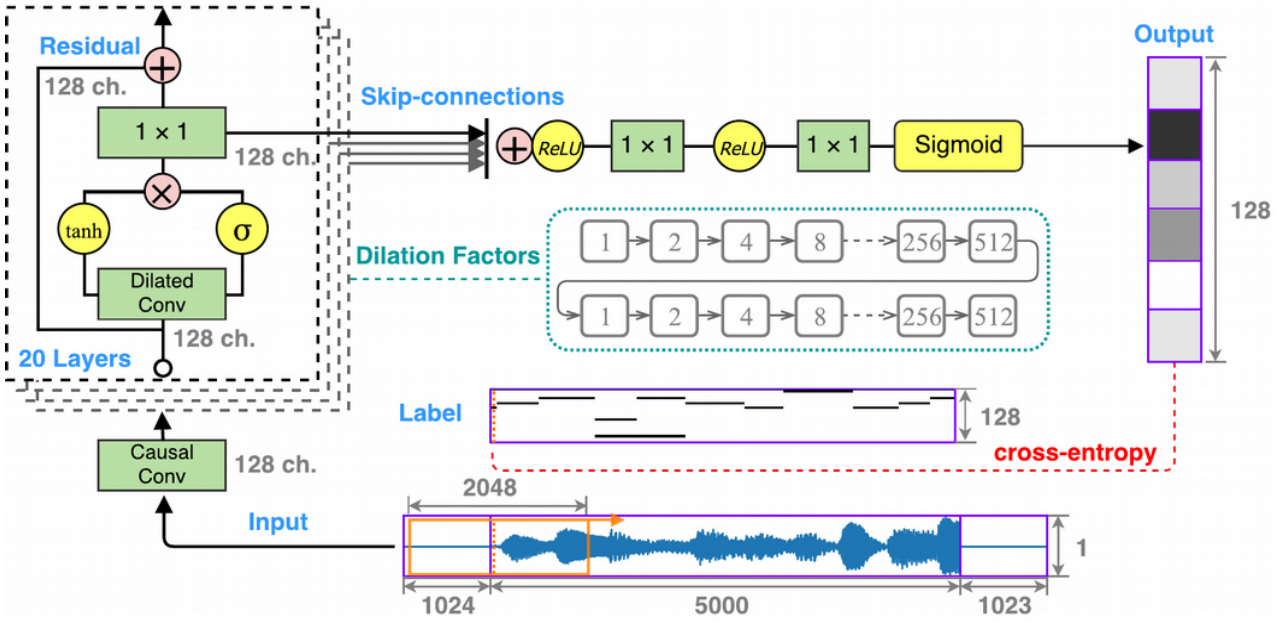
- Конкенативный синтез
- Форматный синтез
- Синтез на основе скрытых марковских моделях
- Глубокое обучение

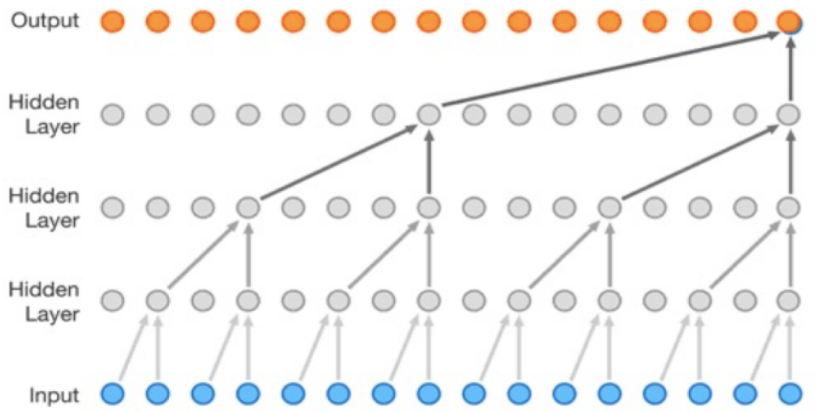
Современные архитектуры

- WaveNet
- DeepVoice
- Tacotron
- DeepVoice 2
- DeepVoice 3
- Tacotron 2

MOS - средняя экспертная оценка разборчивости речи

WaveNet





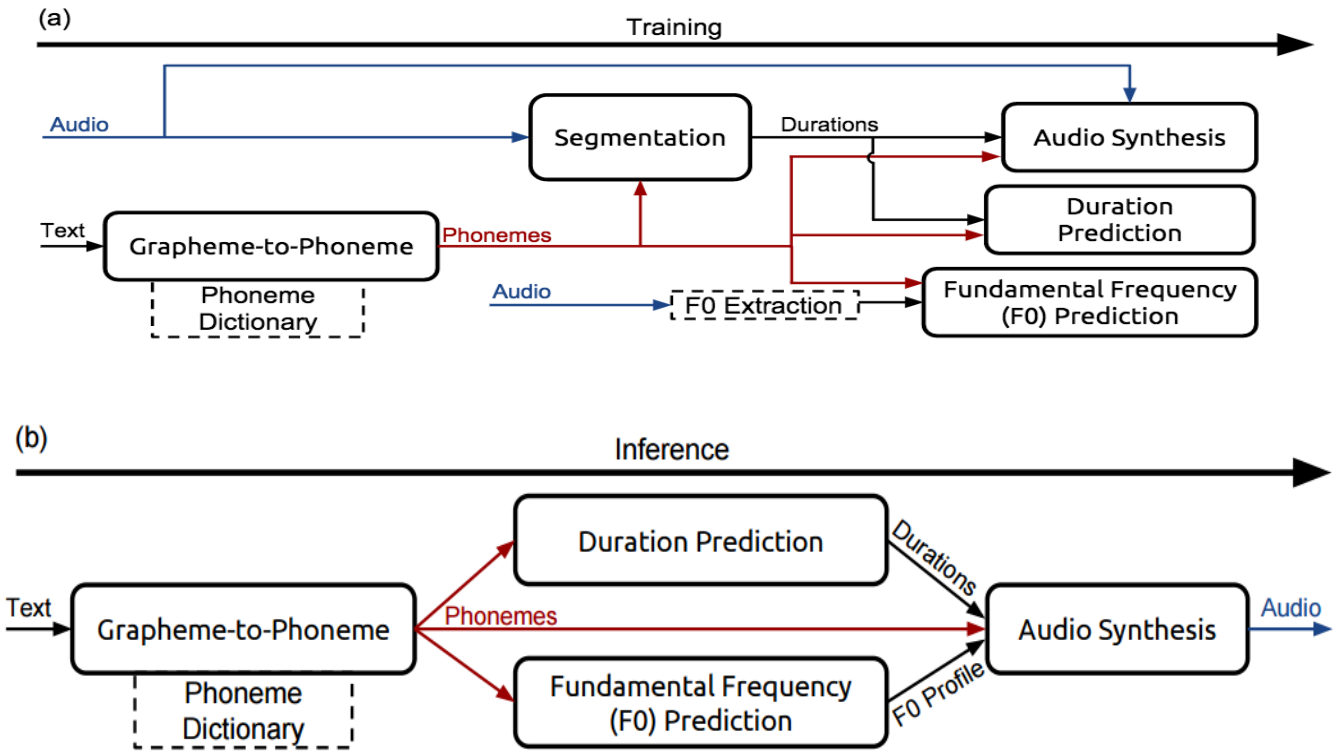
Особенности

- генерирует необработанные звуковые сигналы
- softmax слой моделирует условные распределения по отдельным аудиосемплам
- 4,21 оценка MOS

Минусы

- комплексная система, которая требует подготовки размеченных текстов
- требует дополнительные лингвистические функции (например информацию об ударении или основной частоте)
- вычислительно дорогой синтез

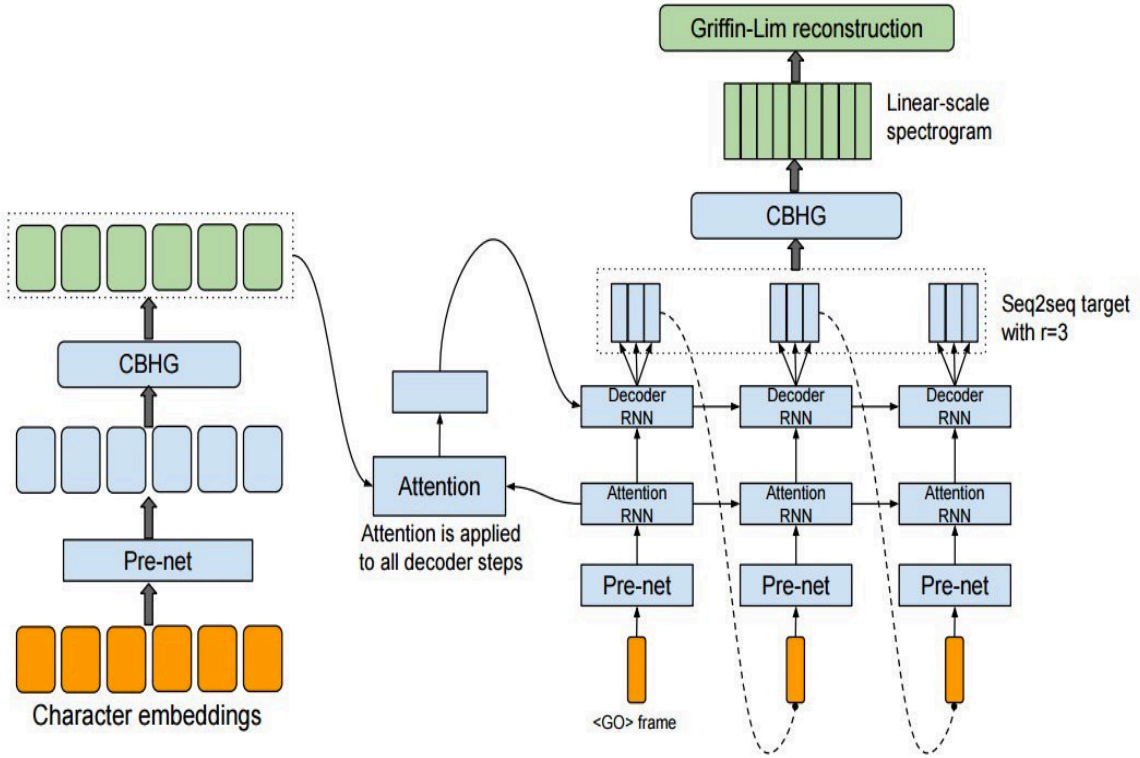
DeepVoice



Особенности

- Закладывает основу для синтеза речи в реальном времени
- 2,67 оценка MOS

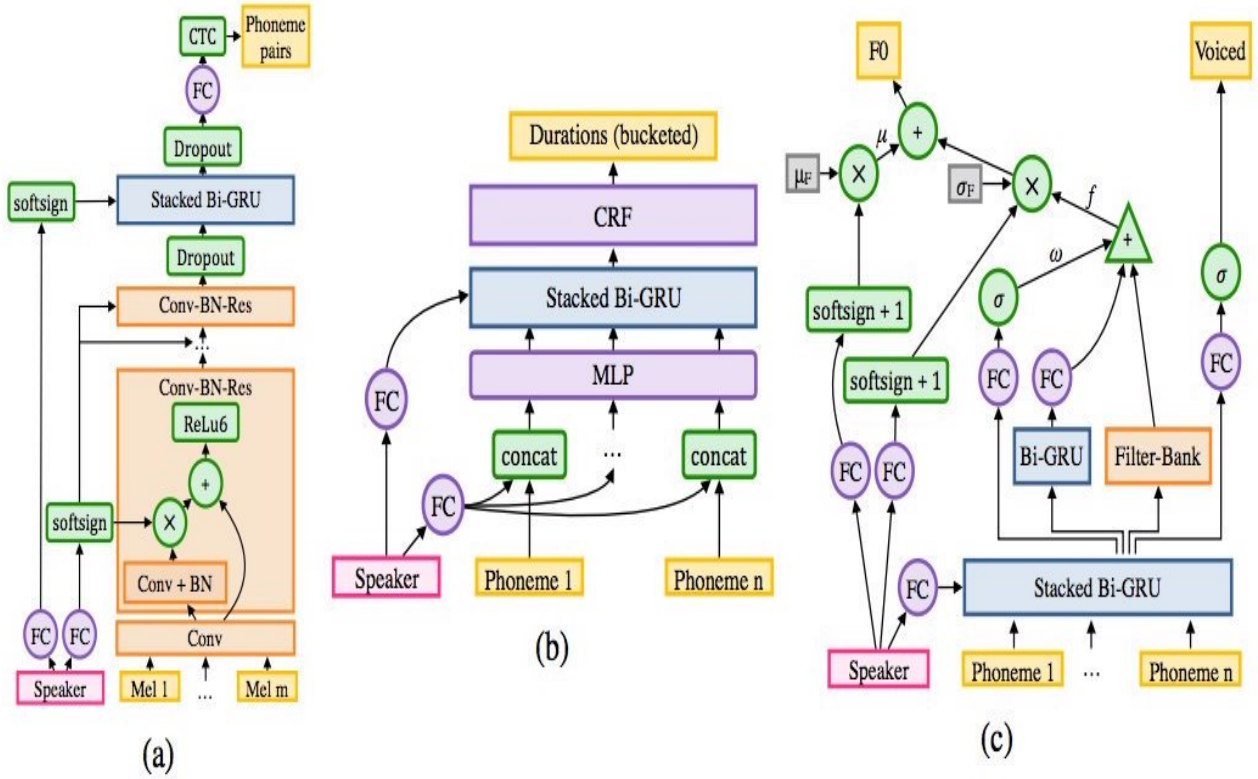
Tacotron



Особенности

- End-to-end модель
- 3,82 оценка MOS

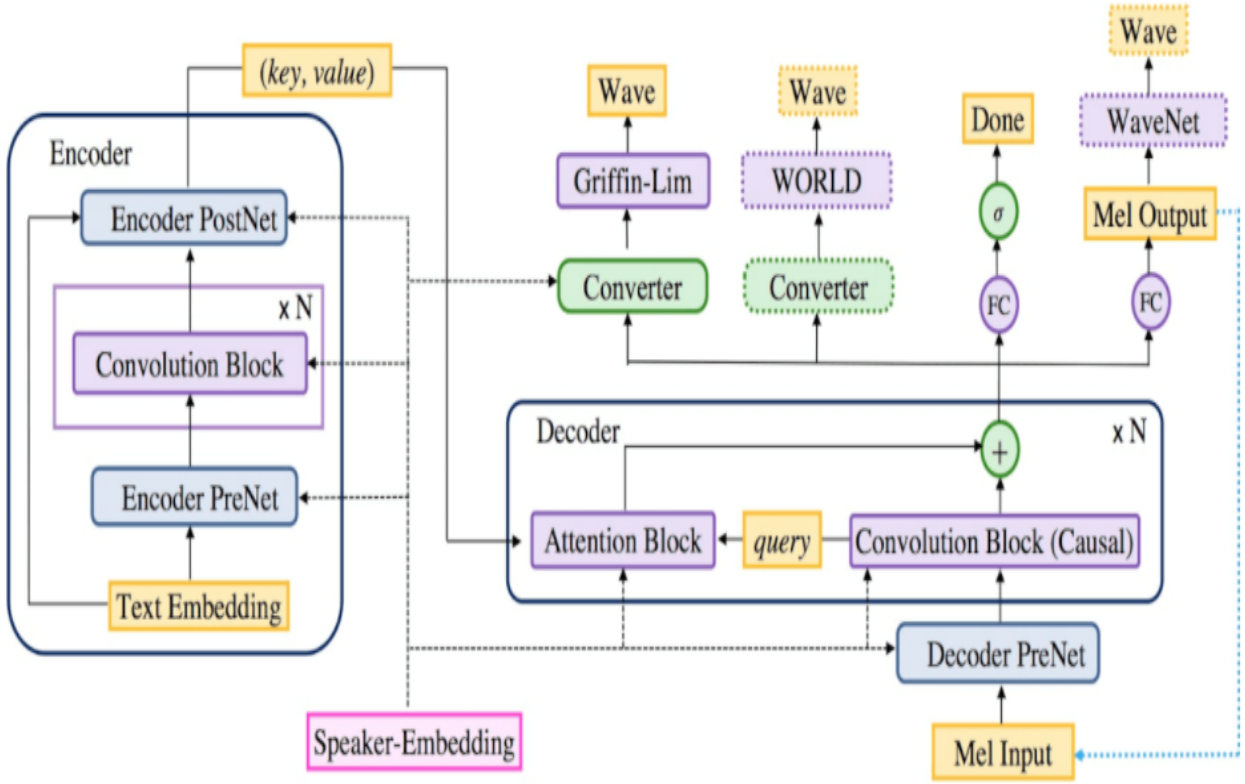
DeepVoice 2



Особенности

- Может выучить множество уникальных голосов
- 3,53 оценка MOS

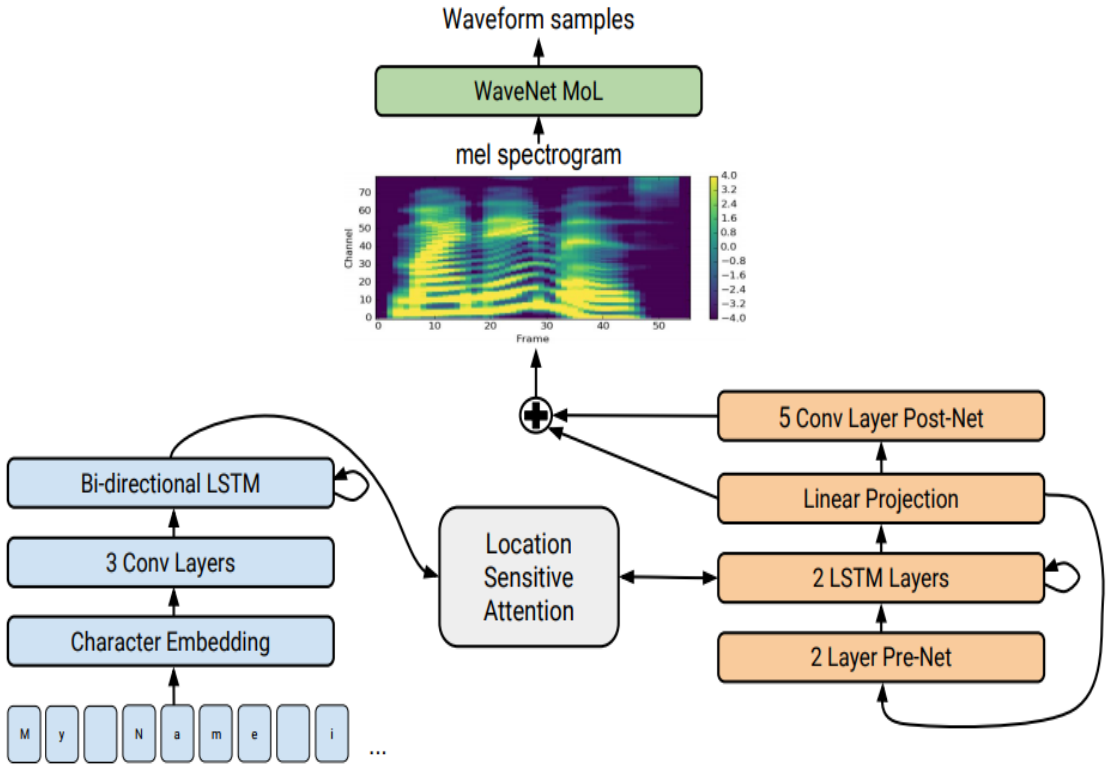
DeepVoice 3



Особенности

- Преобразует входной текст в спектрограмму
- Быстрое обучение
- 3,78 оценка MOS

Tacotron 2



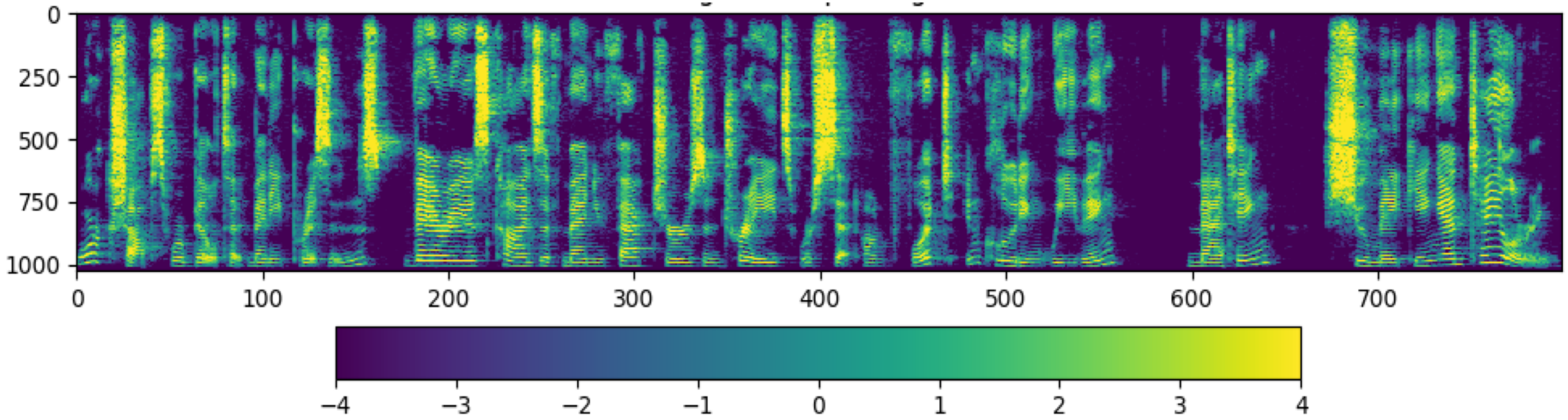
Особенности

- Преобразует входной текст в спектрограмму
- 4,53 оценка MOS

Сравнение архитектур

Архитектура	Открытый исходный код	Возможность скачать пред обученную сеть	Примеры записи	Онлайн демо	MOS
WaveNet	+		+		4,21
DeepVoice	+		+		2,67
Tacotron	+	+	+		3,82
DeepVoice 2	+		+		3,53
DeepVoice 3	+	+	+	+	3,78
Tacotron 2	+		+		4,53

Получение мел-спектрограммы в алгоритме Tacotron 2



Предлагаемое решение

$$\begin{cases} u_l = u_{l-1}(t) * \varphi_l(t) \\ s_l(t) = u_{l-1}(t) * \psi_l(t) \end{cases}$$

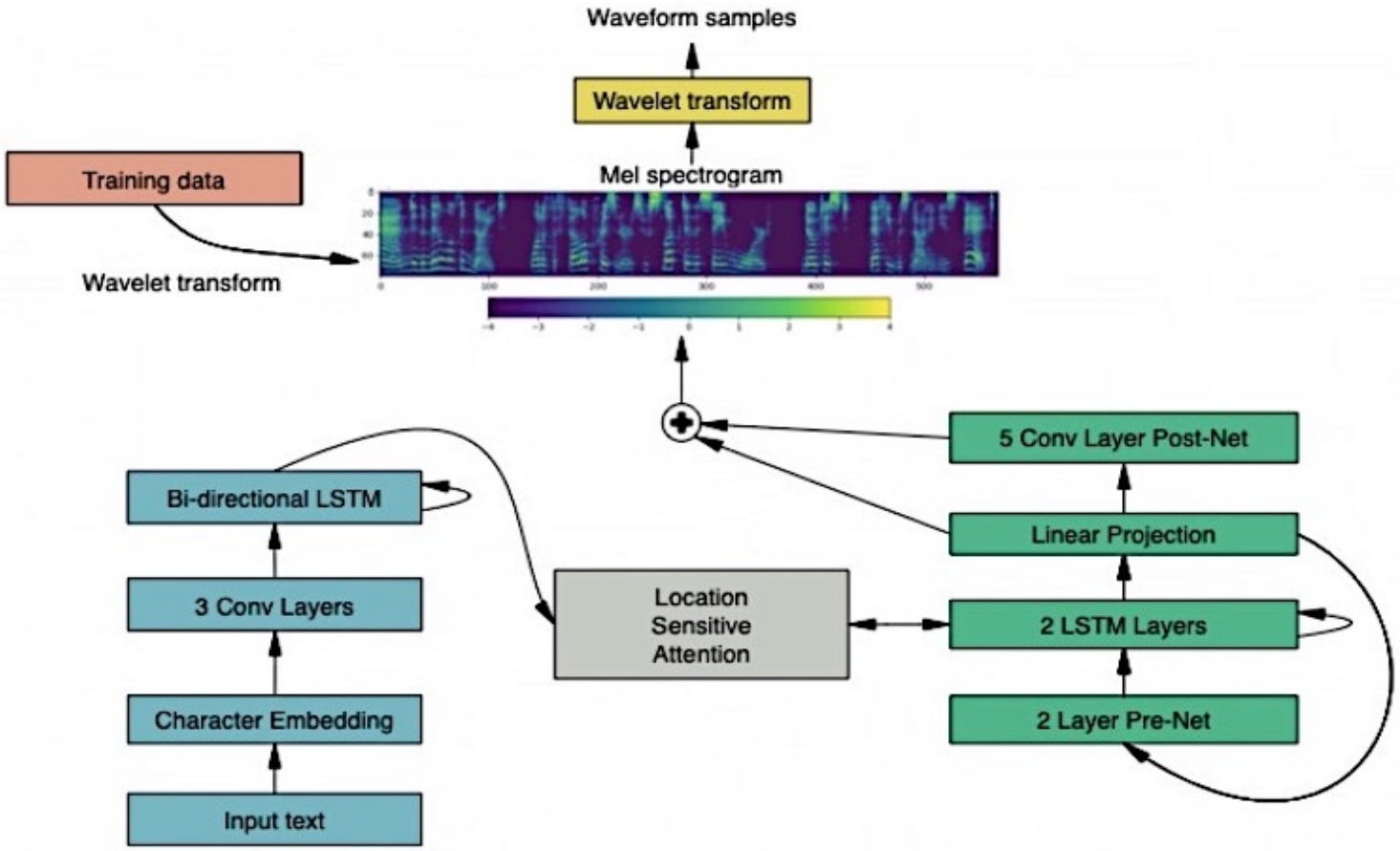
$\varphi_l(t)$ - функции масштабирования Добеши

$\psi_l(t)$ – материнский вейвлет

s_l - поддиапазон звукового сигнала

$u_0 = s(t)$

Разработанный алгоритм



Обучение модели

Набор данных

LJ Speech Dataset: длительность - 23:55:17

Платформа для обучения

Google Colab: TPU - 180 TFLOPS, RAM – 36 Gb

Параметры обучения:

- Метод оптимизации – Adam
- Темп обучения (learning rate) – 0.5
- Изменение темпа обучения (learning rate scheduler) – 40000
- Количество эпох - 40
- Размер батча – 32

Оценка синтезированной речи

PESQ - Воспринимаемая оценка качества речи

$$PESQ = a_0 + a_1 d_{sym} + a_2 d_{asym},$$

где $a_0 = 4{,}5$; $a_1 = -0{,}1$; $a_2 = -0{,}0309$; d_{sym} – среднее значение симметричного возмущения; d_{asym} – среднее значение несимметричного возмущения.

Результаты алгоритма

Алгоритм	**Оценка PESQ**
Wavelet + WaveNet	4.383
Wavelet + Wavelet	3.907
STFT + WaveNet	4.322
STFT + STFT	3.673

Полученные результаты

1. Выполнен аналитический обзор литературы по теме трансляции текста в речь.
2. Рассмотрены современные архитектуры трансляции текста в речь и произведен их сравнительный анализ.
3. Исследована возможность улучшения качества синтезированной речи при помощи вейвлет преобразований.
4. Разработан алгоритм синтеза речи с применением вейвлет преобразования.
5. Реализован разработанный алгоритм и проведено качественное сравнение с существующими решениями

Спасибо за внимание!