

TalkNet: Конволюционная неавторегрессионная модель для задачи генерации речи

Беляев Станислав Валерьевич

Научный руководитель: Гинзбург Б.Е.

Санкт-Петербургская школа физико-математических и компьютерных наук

НИУ ВШЭ – Санкт-Петербург

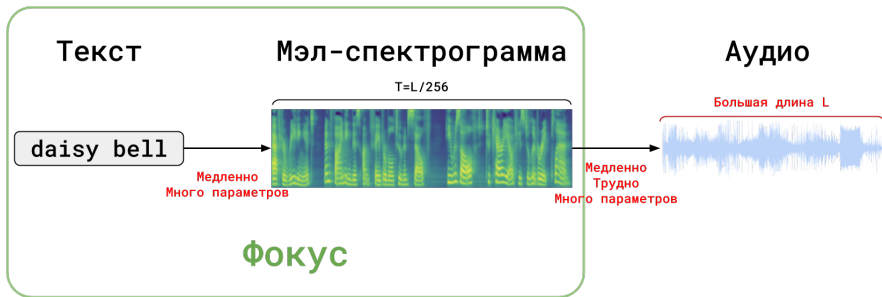
16 июня 2020

Постановка задачи

Задача генерации речи (Text-To-Speech, TTS)

По входному тексту сгенерировать аудиодорожку с человеческой речью.

Обычно разделяют на две части: генерация мэл-спектрограммы (компактное, вычислимое и детерменированное представление аудио) и вокодинг:



Фигура 1: Современный двухшаговый пайплайн генерации речи из текста

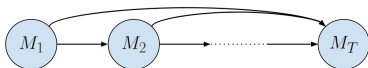
Обзор существующих решений

Устойчивость (robustness)

Способность процесса генерации избегать запинаний и пропуска букв/слов.

Авторегрессионные методы^{1,2}

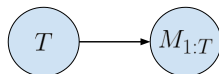
Слева направо, шаг за шагом



- Медленное обучение и генерация
- Отсутствие устойчивости
- Много обучаемых параметров
- + Хорошее качество

Неавторегрессионные методы³

Параллельно



- + Быстрое обучение и генерация
- + Устойчивость
- Много обучаемых параметров
- Чуть хуже качество

¹J. Shen et al. "Natural TTS Synthesis by Conditioning Wavenet". In: *ICASSP*. 2018.

²Naihan Li et al. "Neural Speech Synthesis with Transformer Network". In: *AAAI*. 2019.

³Yi Ren et al. "FastSpeech: Fast, Robust and Controllable Text to Speech". In: *NeurIPS*. 2019.

Цели и задачи

Цель

Целью данной работы является разработка нового подхода для задачи синтеза речи, позволяющего производить **быструю, качественную и устойчивую** генерацию с **малым** количеством обучаемых параметров.

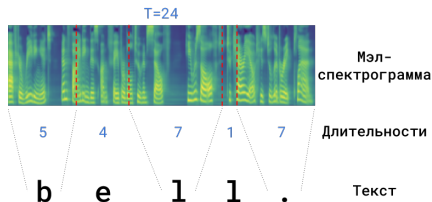
Задачи:

- Разработать и описать эффективную нейронную архитектуру, основанную на идеи неавторегрессионности.
- Выбрать данные для обучения, провести эксперименты и замерить результаты.
- Произвести сравнение качества и устойчивости генерации с существующими решениями.
- Проанализировать скорость генерации и сравнить с другими подходами.

Идея

Длительность символа

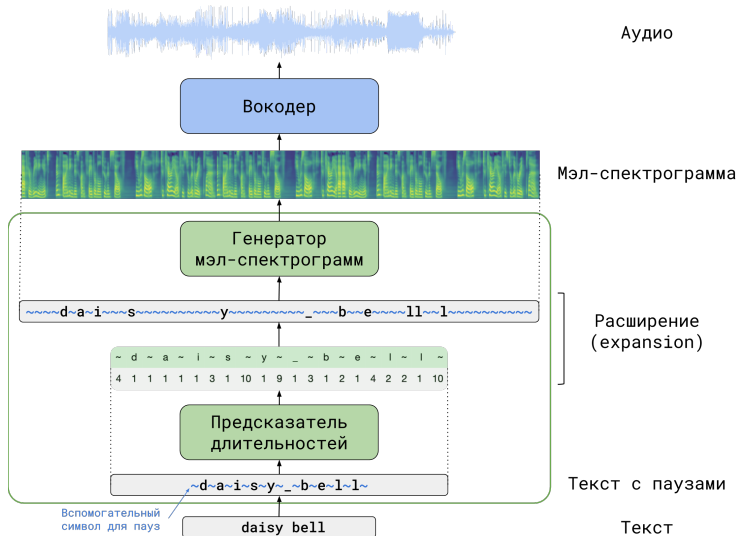
Количество шагов в спектрограмме, соответствующих выводу символа.



Фигура 2: Временное выравнивание текста и аудио

- Зная истинные длительности всех символов, можно построить **неавторегрессионный** генератор мэл-спектрограммы. Хорошее выравнивание позволяет получить свойство **устойчивости**.
- Длительности датасета можно извлечь из модели, решающей обратную задачу – распознавания речи (Automatic Speech Recognition, ASR).
- Этап предсказания длительностей для символов входного текста требует отдельной **неавторегрессионной** нейронной модели.

Архитектура



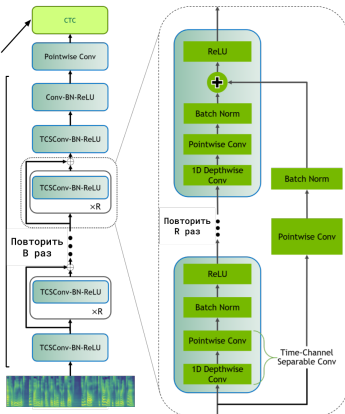
Фигура 3: Архитектура генерации аудио TalkNet

Извлечение истинных длительностей

Функция ошибки, явным образом учащая выравнивание спектрограммы и текста с паузами по времени

Полносверточная архитектура
QuartzNet BxR

Мэл-спектрограмма



ASR легче TTS:
средняя ошибка CTC =
~1 неправильно
распознанный символ
для датасета в
1000ч. аудио

Фигура 4: QuartzNet⁴ решает обратную задачу: распознавания текста по аудио. Функция ошибки учит выравнивание, из которого извлекаются длительности букв.

⁴Samuel Krman, Stanislav Beliaev, and Boris Ginsburg. "QuartzNet: deep automatic speech recognition with 1D time-channel separable convolutions". In: *ICASSP*. 2020.

Процесс обучения

Этап	Архитектура	Количество параметров	Время обучения
Предсказатель длительностей	QuartzNet 5x5	2M	15м.
Генератор спектрограммы	QuartzNet 9x5	8M	2ч.

Таблица 1: Параметры обучения для двух независимых этапов TalkNet

- Для обучения использовался набор данных LJSpeech⁵ – 24 часа одноголосой речи из аудиокниг с художественной литературой.
- Общее количество обучаемых параметров модели – **10M** – в **3** раза меньше чем у любого другого современного подхода.
- Общее количество времени на обучение – **2 часа** – в **20** раз быстрее лучших авторегрессионных методов.
- Для обучения применялся случайный несмешенный шум для истинных длительностей, что стабилизировало процесс и повысило качество.

⁵Keith Ito et al. *The LJ speech dataset*. 2017.

Сравнение качества

Mean Opinion Score (MOS)

Слепая оценка носителей языка натуральности речи предложенного аудио по дискретной шкале от 1.0 до 5.0 с шагом 0.5. Каждый пример оценивается несколькими разными слушателями для понижения дисперсии.

Подход	MOS	} Небольшая разница
Оригинальное аудио	4.31 ± 0.05	
Оригинальный мэл + WaveGlow	4.04 ± 0.05	
Tacotron 2 + WaveGlow	3.85 ± 0.06	
TalkNet + WaveGlow	3.74 ± 0.07	

Таблица 2: Оценки MOS с использованием вокодера WaveGlow и 95% доверительным интервалом. Оценки усреднялись по 100 примерам, каждый из которых был оценен не менее 10 раз. Как можно заметить, **качество TalkNet сравнимо с Tacotron 2** – авторегрессионным подходом, считающихся одним из лучших на сегодняшний момент.

Сравнение скорости

Latency

Количество секунд, требуемых на генерацию одного аудио.

Real Time Factor (RTF)

Количество секунд аудио, генерирующихся за одну секунду вычислений.

Model	Latency, c	RTF
Transformer TTS	6.735 ± 3.969	1.48 ± 0.87
Tacotron 2	$0.817 \pm 1 \cdot 10^{-2}$	7.56 ± 0.01
FastSpeech	$0.029 \pm 2 \cdot 10^{-4}$	221.01 ± 1.75
TalkNet	$0.019 \pm 1 \cdot 10^{-5}$	328.65 ± 4.76

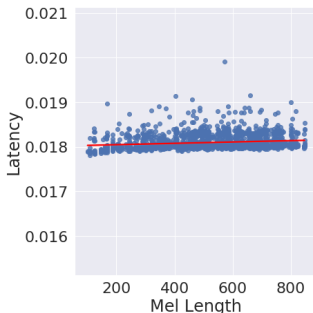
Таблица 3: Сравнение скорости генерации с 95% доверительным интервалом. Количество примеров для усреднения – 2048, а средняя длина мэл-спектрограммы – 540. Как видно, **TalkNet значительно быстрее** современных методов.

Генерация за $O(1)$

Механизм внимания (attention)

Вычислительная операция для упрощения поиска взаимосвязей в данных, работающая квадратично от времени аудио (длины спектрограммы).

Используется **во всех** современных моделях для генерации речи.



Фигура 5: График зависимости длины мела от latency. Неавторегрессионность и **отсутствие** механизмов внимания позволяют получить **константное** время вывода.







Результаты

- Разработана **устойчивая** неавторегрессионная модель для синтеза речи, которая использует предобученную модель задачи распознавания речи для извлечения истинных длительностей символов.
- Подход обладает **высокой** скоростью обучения (до **20** раз в сравнении другими моделями) и генерации (в **320** раз быстрее реального времени). Скорость генерации **не зависит** от длины входного текста.
- Архитектура подхода позволяет **уменьшить** количество обучаемых параметров до **10M** – в **3** раза меньше по сравнению с конкурентами, показывая при этом **сравнимое** качество.
- Статья⁶ (INTERSPEECH 2020)
- Код⁷

⁶Stanislav Beliaev, Yurii Rebryk, and Boris Ginsburg. *TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model*. 2020. [arXiv: 2005.05514 \[eess.AS\]](https://arxiv.org/abs/2005.05514).

⁷<https://github.com/NVIDIA/NeMo/tree/master/examples/tts>

Литература

-  Beliaev, Stanislav, Yurii Rebyrk, and Boris Ginsburg. *TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model*. 2020. arXiv: 2005.05514 [eess.AS].
-  Ito, Keith et al. *The LJ speech dataset*. 2017.
-  Krیمان, Samuel, Stanislav Beliaev, and Boris Ginsburg. "QuartzNet: deep automatic speech recognition with 1D time-channel separable convolutions". In: *ICASSP*. 2020.
-  Li, Naihan et al. "Neural Speech Synthesis with Transformer Network". In: *AAAI*. 2019.
-  Ren, Yi et al. "FastSpeech: Fast, Robust and Controllable Text to Speech". In: *NeurIPS*. 2019.
-  Shen, J. et al. "Natural TTS Synthesis by Conditioning Wavenet". In: *ICASSP*. 2018.