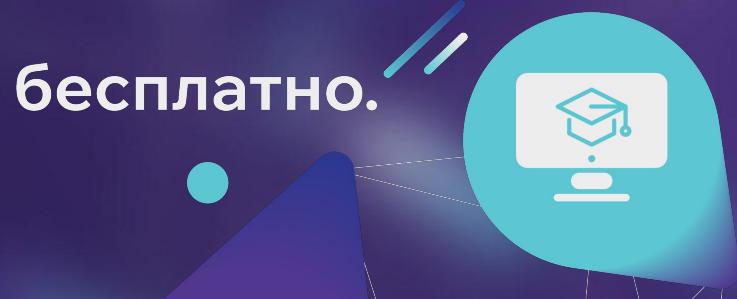


# Deep Learning School

бесплатно.



онлайн.



фундаментально.

2024

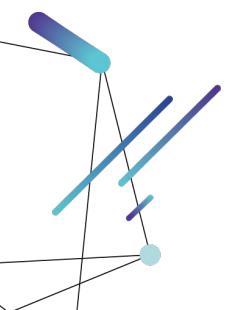
# Генерация речи. Лекция 1

Садекова Таснима

Huawei

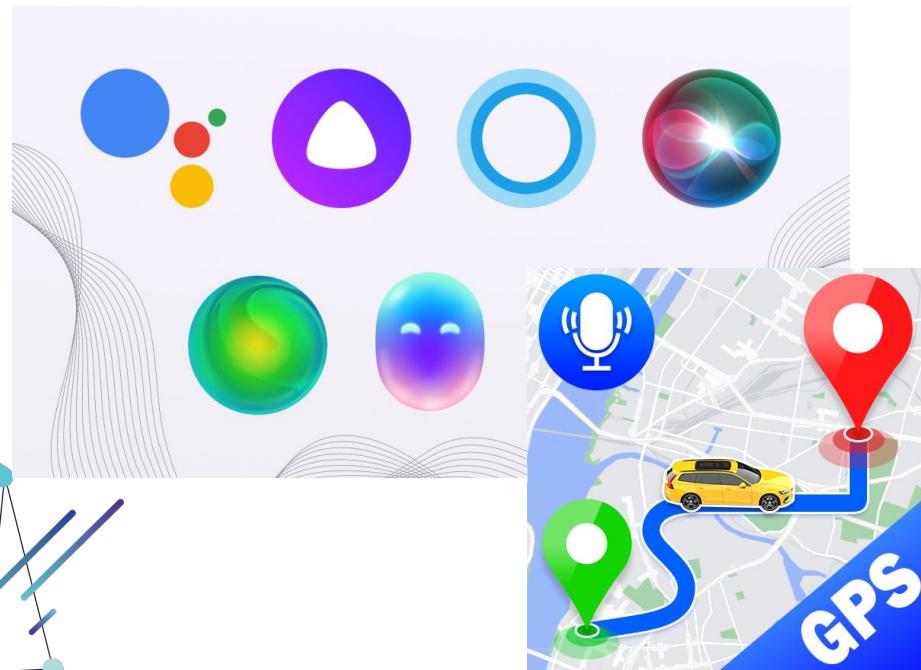
# Содержание

1. Применение
2. История
3. TTS pipeline
4. Метрики качества
5. Акустические модели
6. Вокодеры



# Применение

Голосовые ассистенты  
Интуитивные интерфейсы



Аудиокниги  
Озвучивание образовательных  
материалов



Озвучивание персонажей  
фильмов, игр



Speech-to-speech translation  
Дубляж видео



# История. Машина фон Кемпелена

1788

Машина фон Кемпелена – прототип голосового тракта человека

поток воздуха – воздушные меха

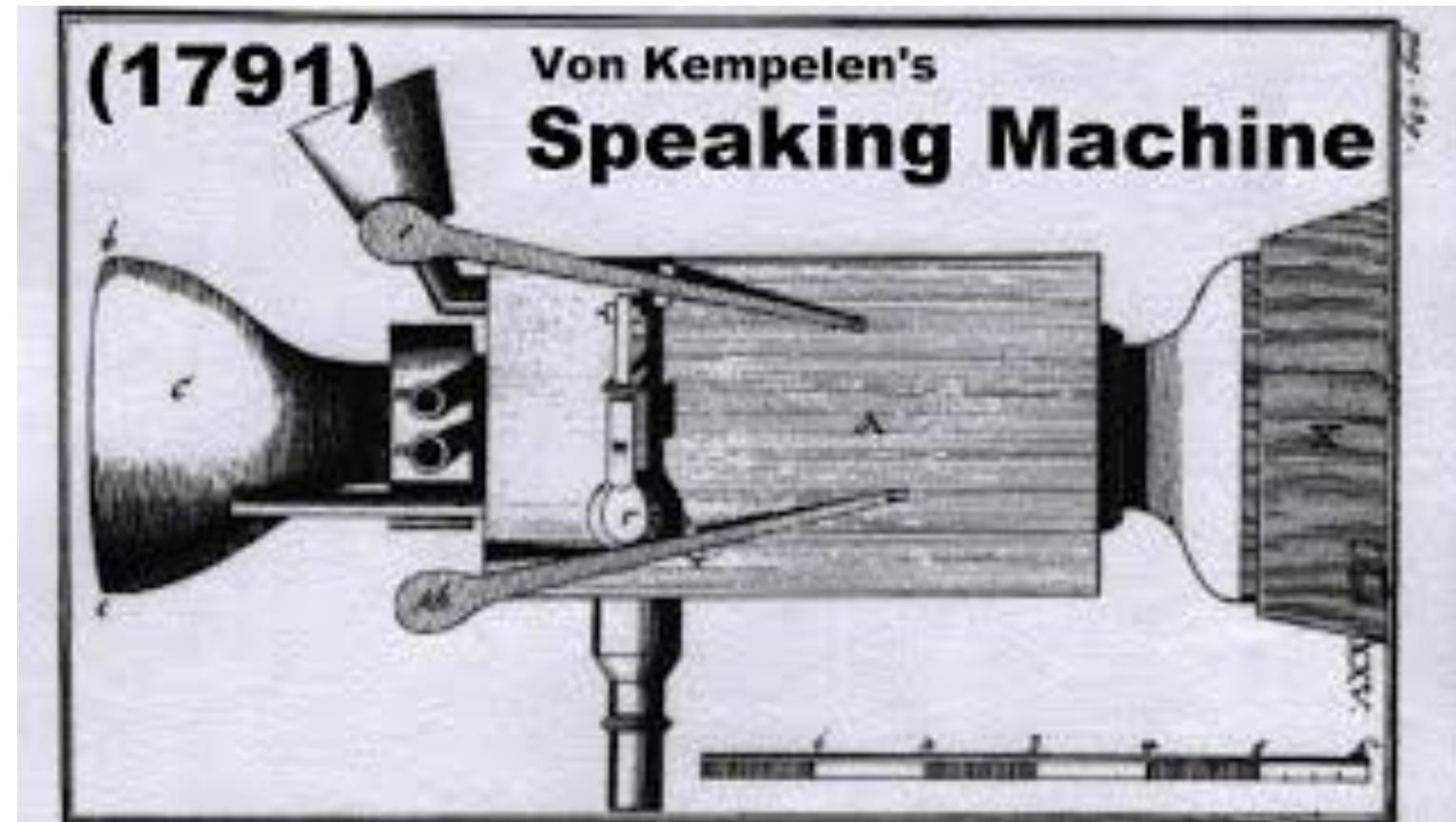
положение голосового тракта – трубы разной ширины

Ряд гласных и согласных звуков



# История. Машина фон Кемпелена

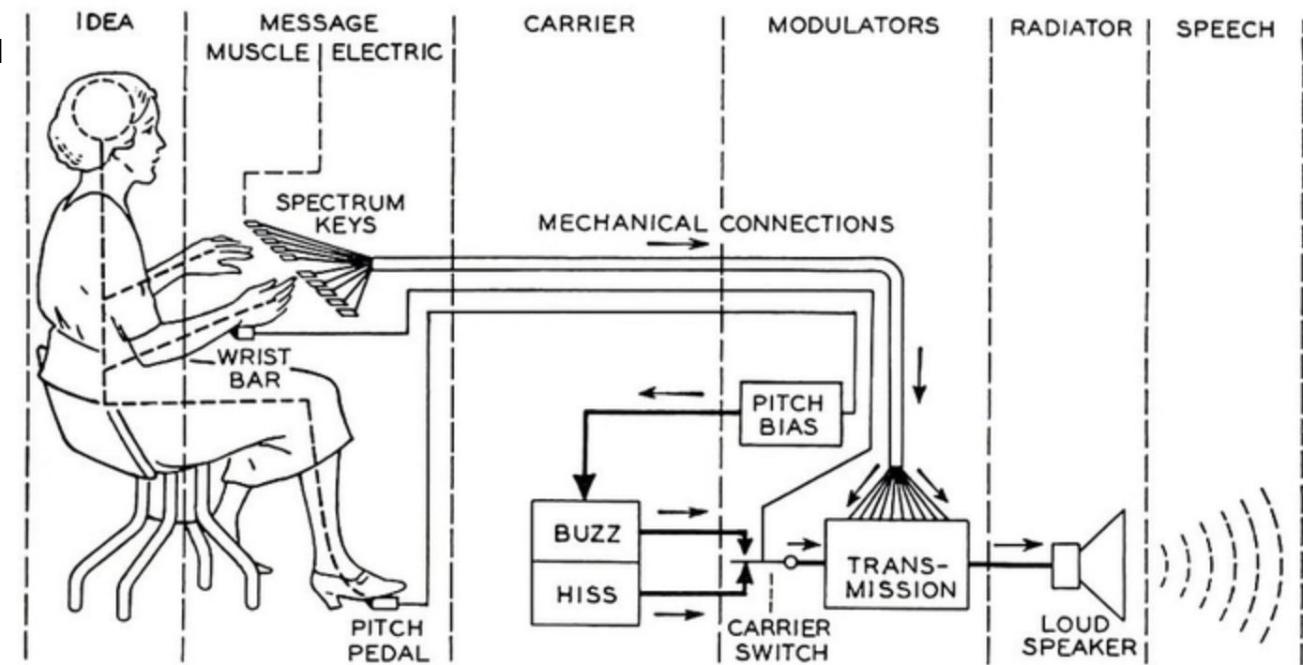
1788



# История. Voder

1938

- **Voder** (voice operating demonstrator) – первый синтезатор речи, предложенный Гомером Дадли из Bell Labs.
- Также аналог голосового тракта человека
- Состоял из:
  - педали
  - 10 клавиш-фильтров
  - рычага, управляемого запястьем
- Синтез целых предложений
- Можно было добавлять интонацию, менять голос
- Звук “роботизированный”



# История. Voder

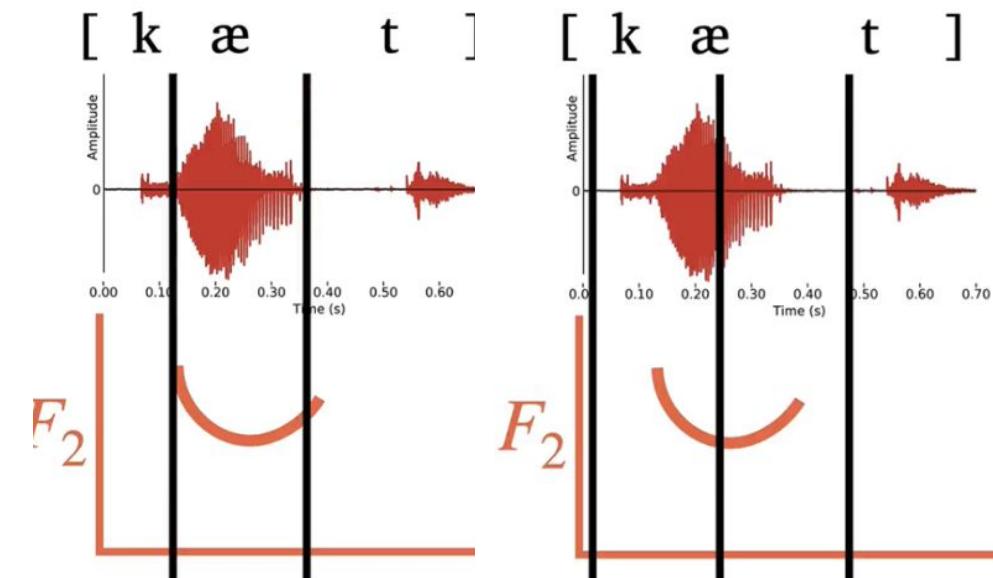
1938



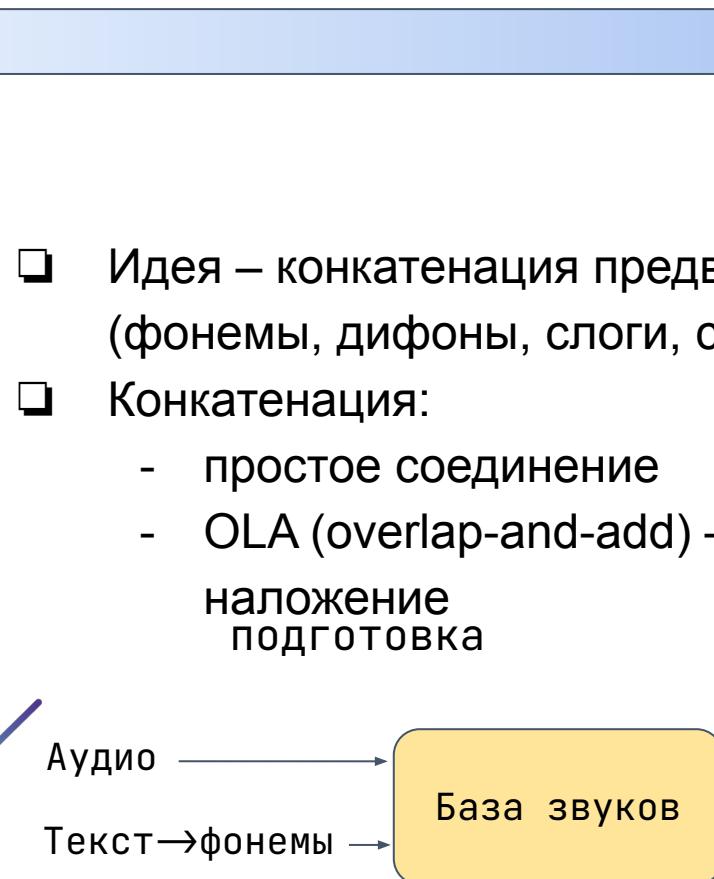
# История. Конкатенативный синтез



- Идея – конкатенация предварительно записанных звуков (фонемы, дифоны, слоги, слова, ...)
- Конкатенация:
  - простое соединение
  - OLA (overlap-and-add) – затухание на краях и наложение
- [Интерактивное демо](#)



# История. Конкатенативный синтез



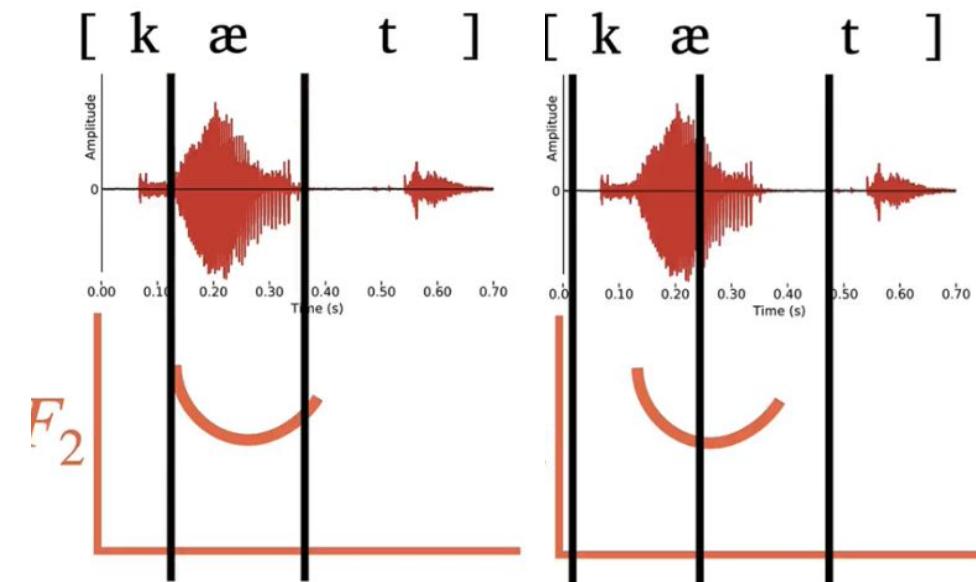
- ❑ Идея – конкатенация предварительно записанных звуков (фонемы, дифоны, слоги, слова, ...)

- ❑ Конкатенация:
  - простое соединение
  - OLA (overlap-and-add) – затухание на краях и наложение подготовка

1970

генерация

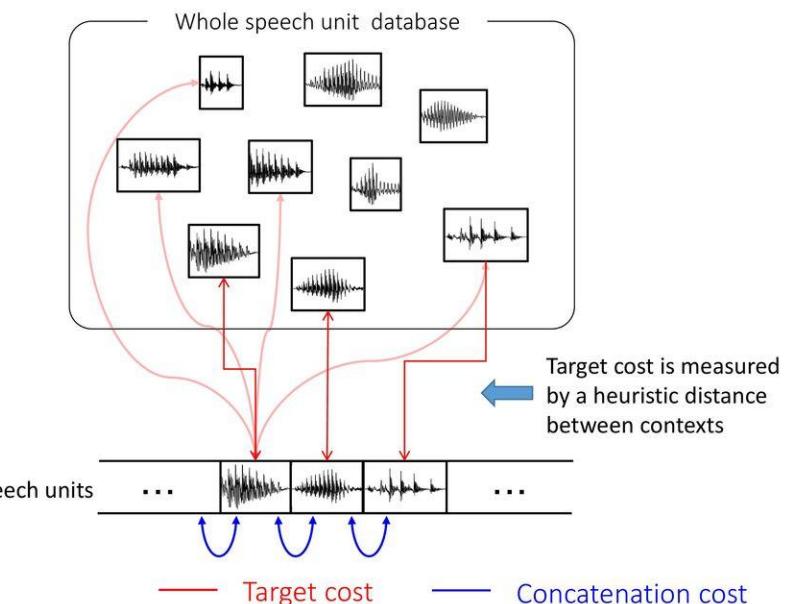
Текст→фонемы



# История. Конкатенативный синтез



Concatenative synthesis



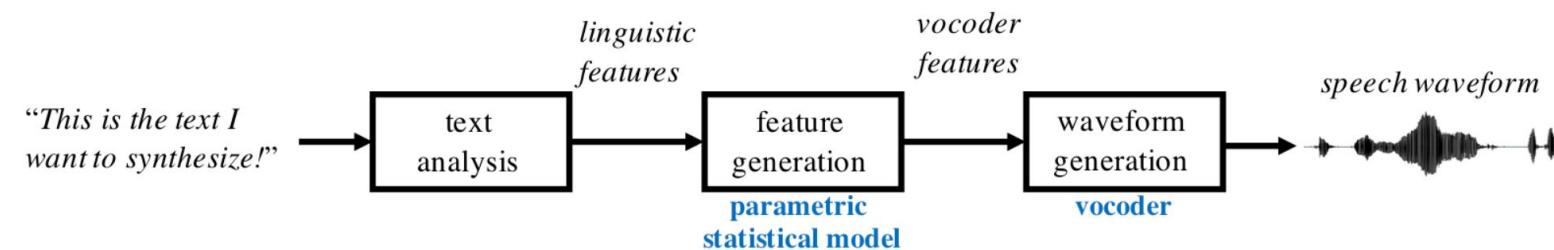
## Unit selection

- Несколько записей на один и тот же юнит. Более подробные описания
  - /ih-t/, +stress, phrase internal, high F0, content word
  - /n-t/, -stress, phrase final, high F0, function word
  - /dh-ax/, -stress, phrase initial, low F0, word 'the'
- Конкатенация: TD-PSOLA (time domain pitch synchronous OLA)
- Выбор из базы с помощью алгоритма Витерби + beam search
- Использовался как первое решение для голосовых помощников Siri, Алиса

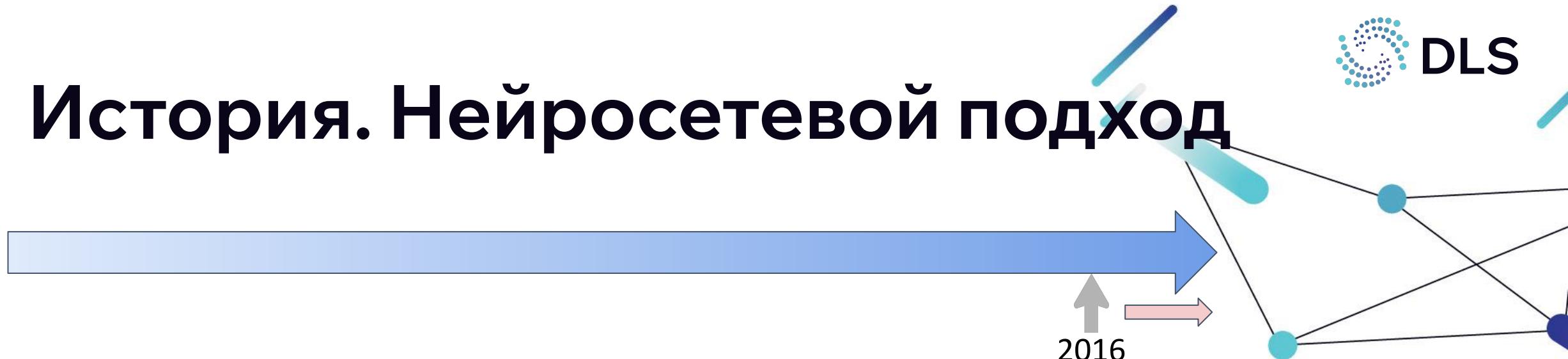
# История. Параметрический синтез



- ❑ Аудио – набор характеристик-параметров (e.g. spectral envelope, F0, periodicity characteristics)
- ❑ Два блока:
  - 1) Статистическая параметрическая модель: по тексту предсказывает параметры  
Hidden-Markov model Gaussian mixture model (HMM-GMM)
  - 2) Вокодер: по параметрам предсказывает звуковую волну (e.g STRAIGHT)
- ❑ Использовался в Amazon Alexa, Google Assistant



# История. Нейросетевой подход



- ❑ Акустическая модель и вокодер – нейронные сети
- ❑ Все промежуточные параметры заменяются на мел-спектrogramмы

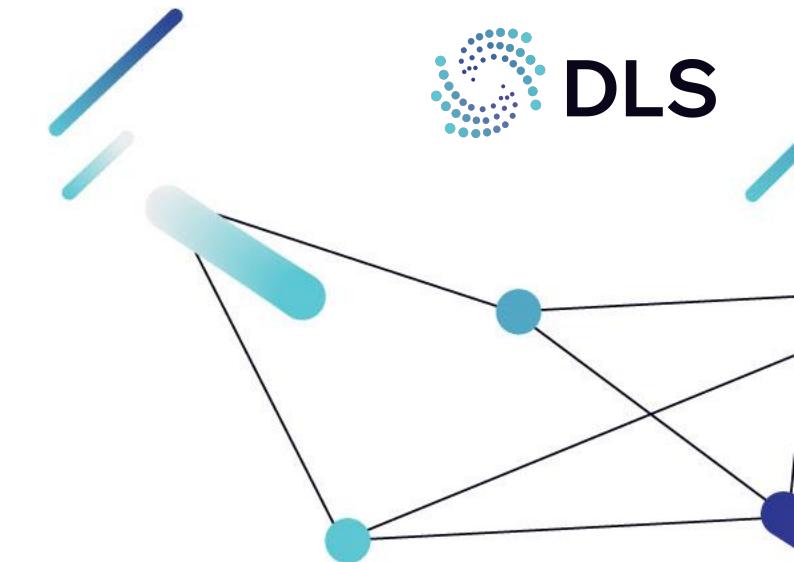
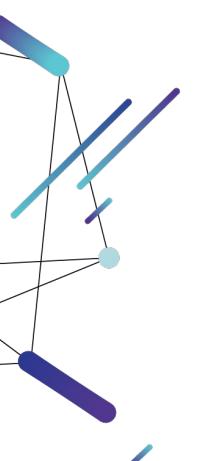


# TTS pipeline

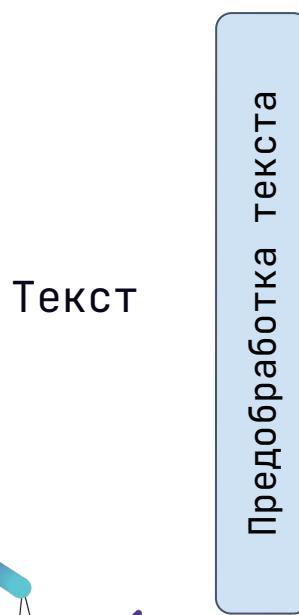
# TTS pipeline

Текст

Аудио

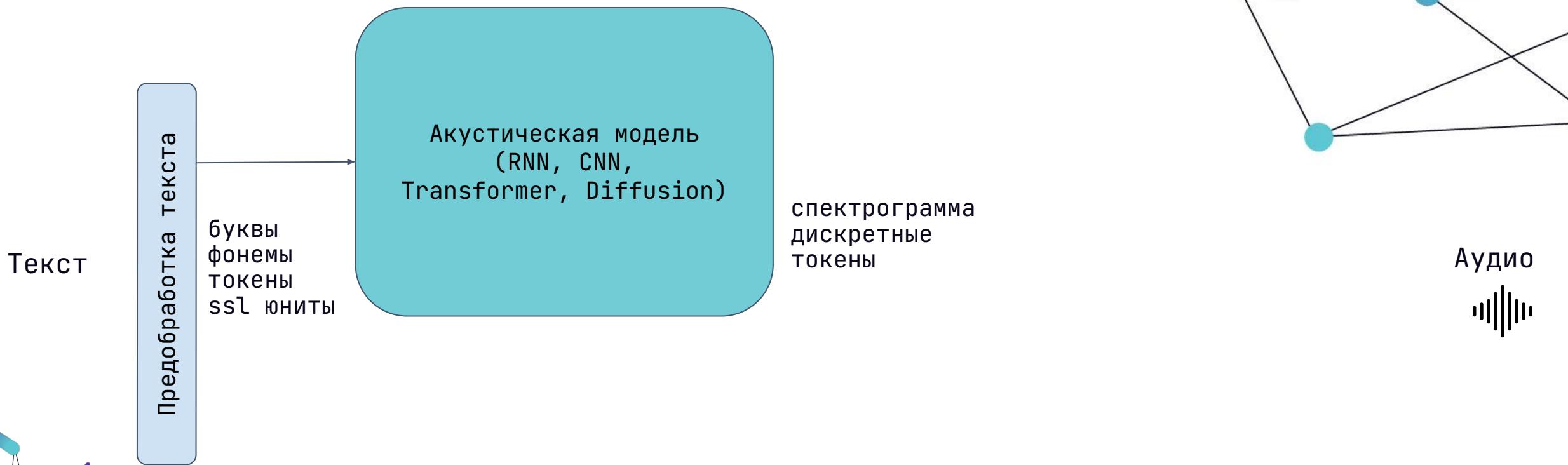


# TTS pipeline

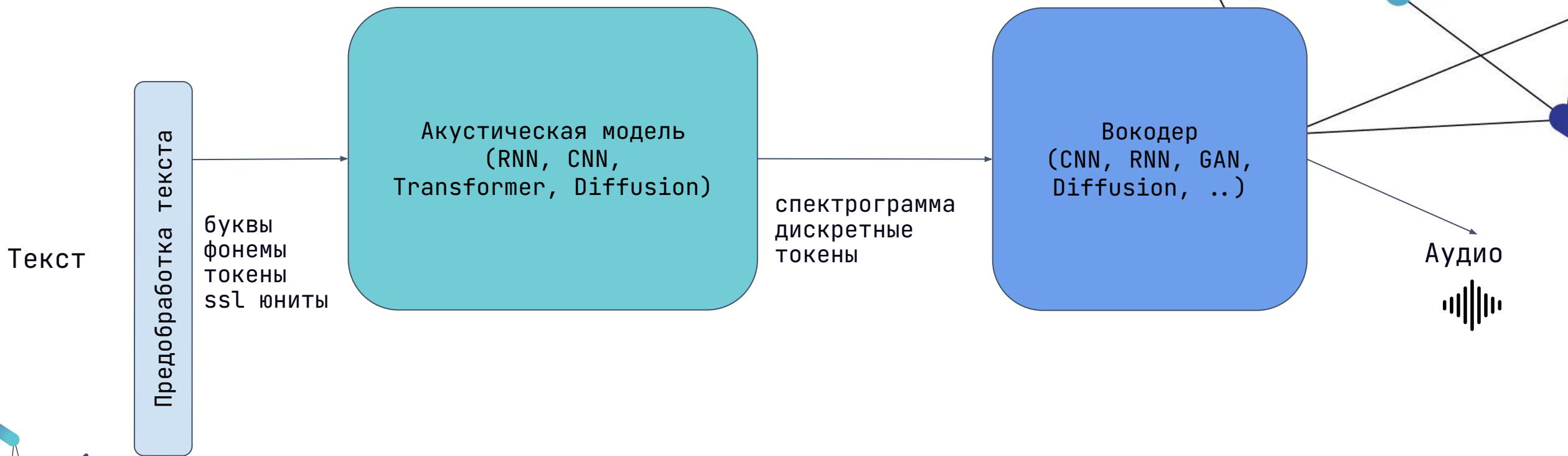


1. Нормализация текста
  - числа, специальные символы, сокращений -> слова
  - пунктуация
1. Применение фонемизатора/токенизатора/ssl модели

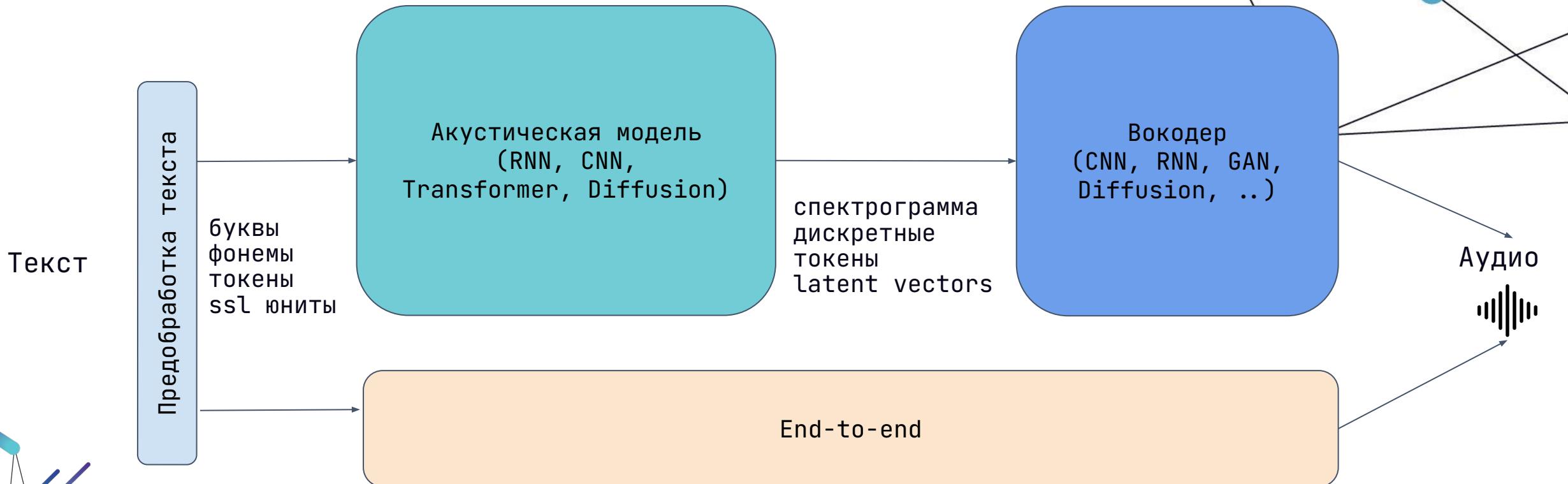
# TTS pipeline



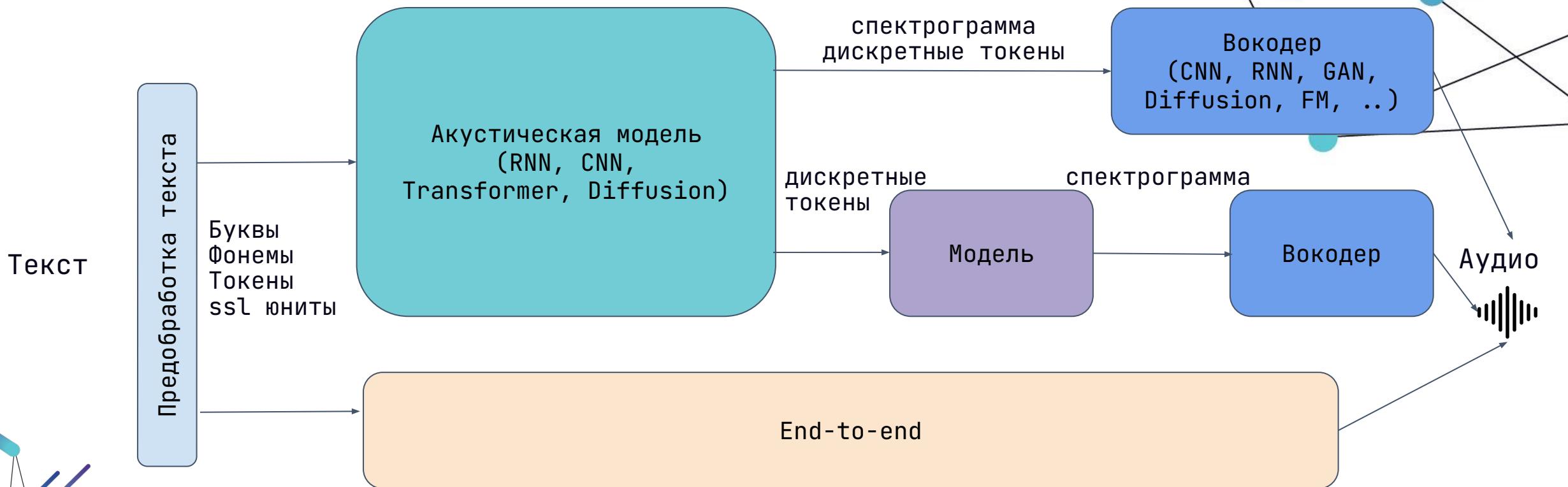
# TTS pipeline



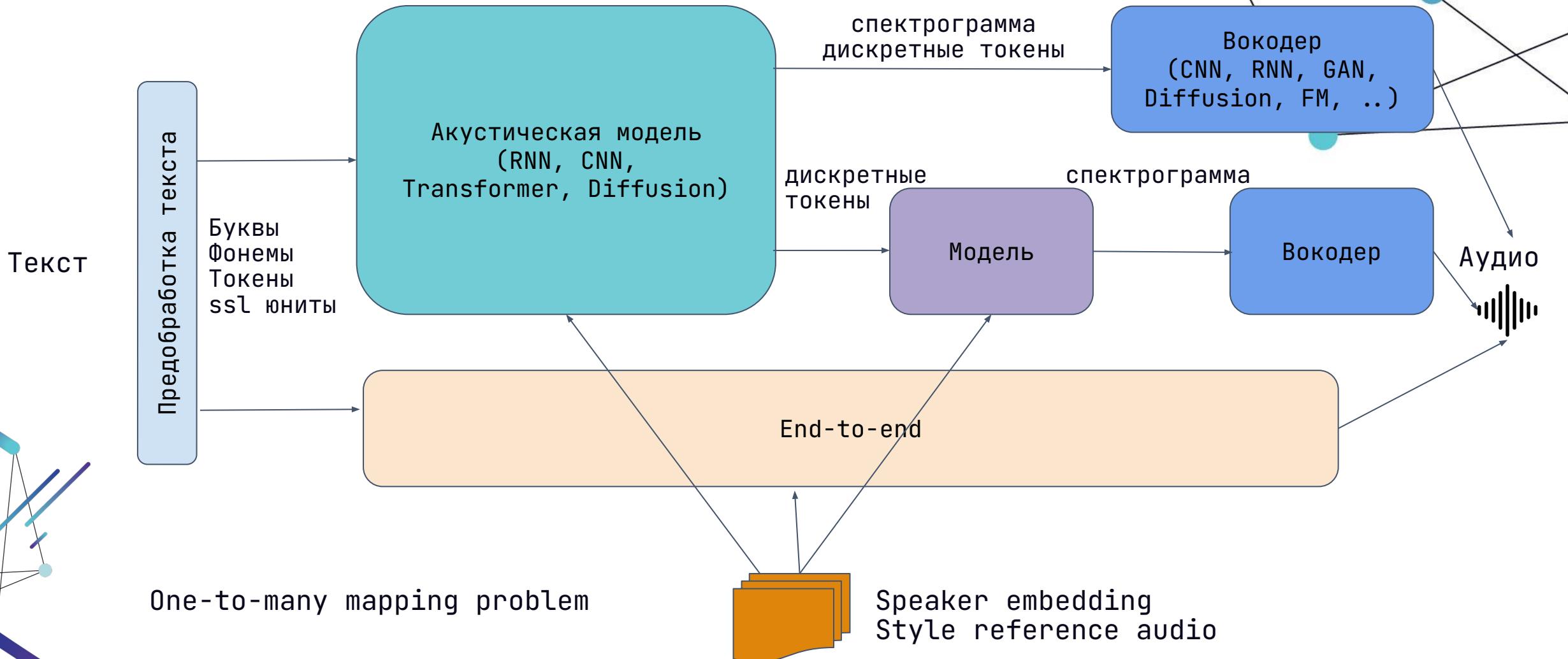
# TTS pipeline



# TTS pipeline



# TTS pipeline



# Близкие задачи

**Voice cloning**

Text →

Voice cloning  
acoustic model



Vocoder

New speaker's  
voice

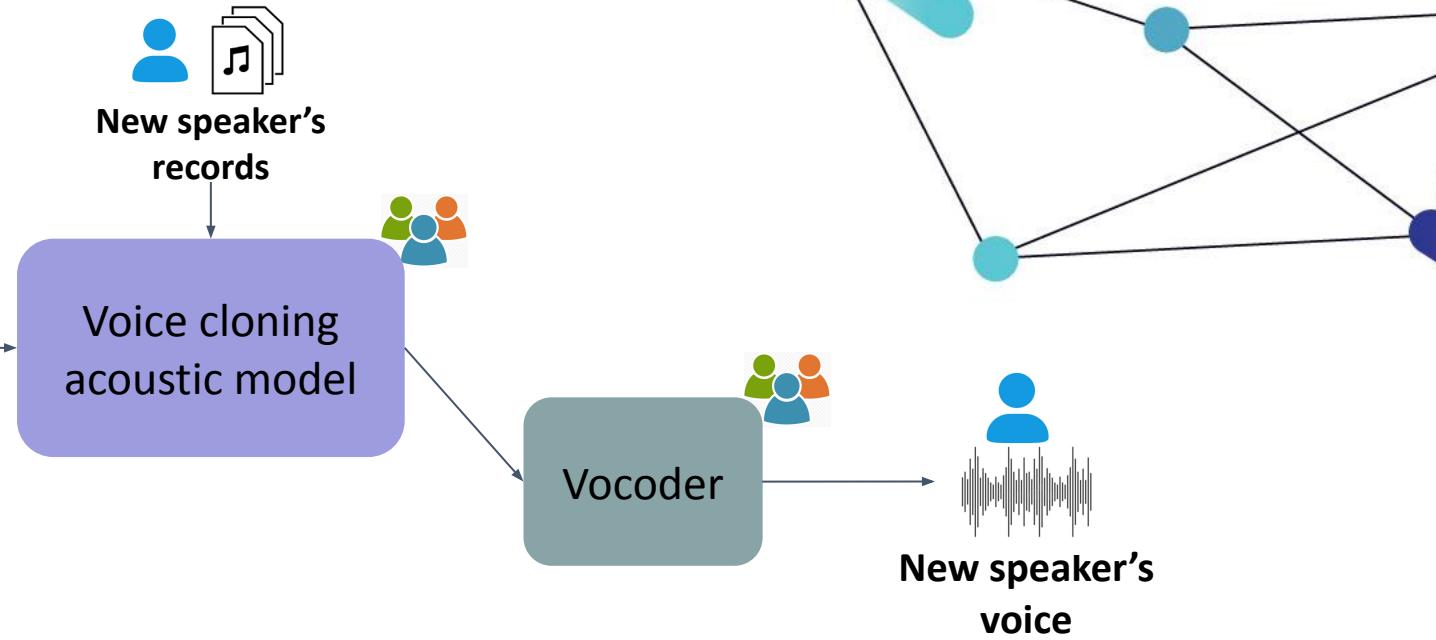
**Voice conversion**

# Близкие задачи

## Voice cloning

- zero-shot
- few-shot

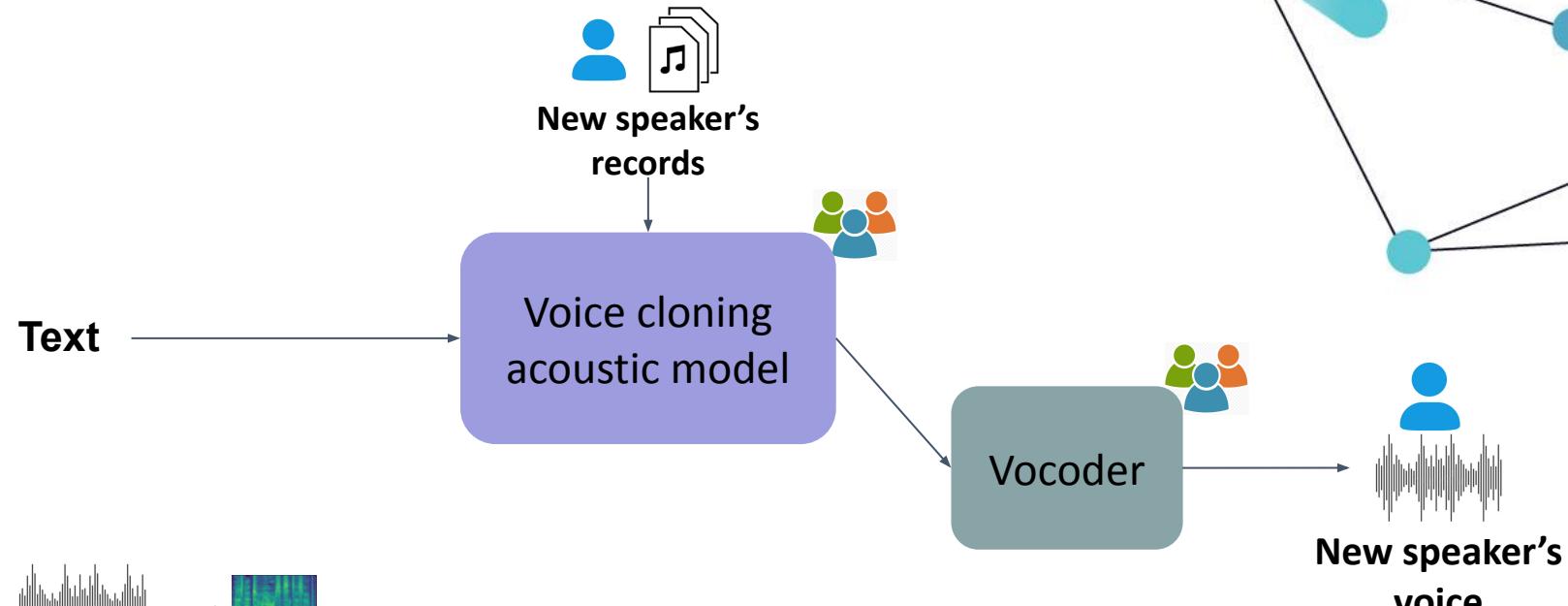
Text →



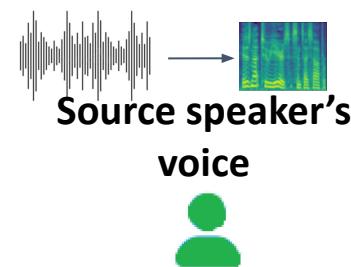
## Voice conversion

# Близкие задачи

Voice cloning

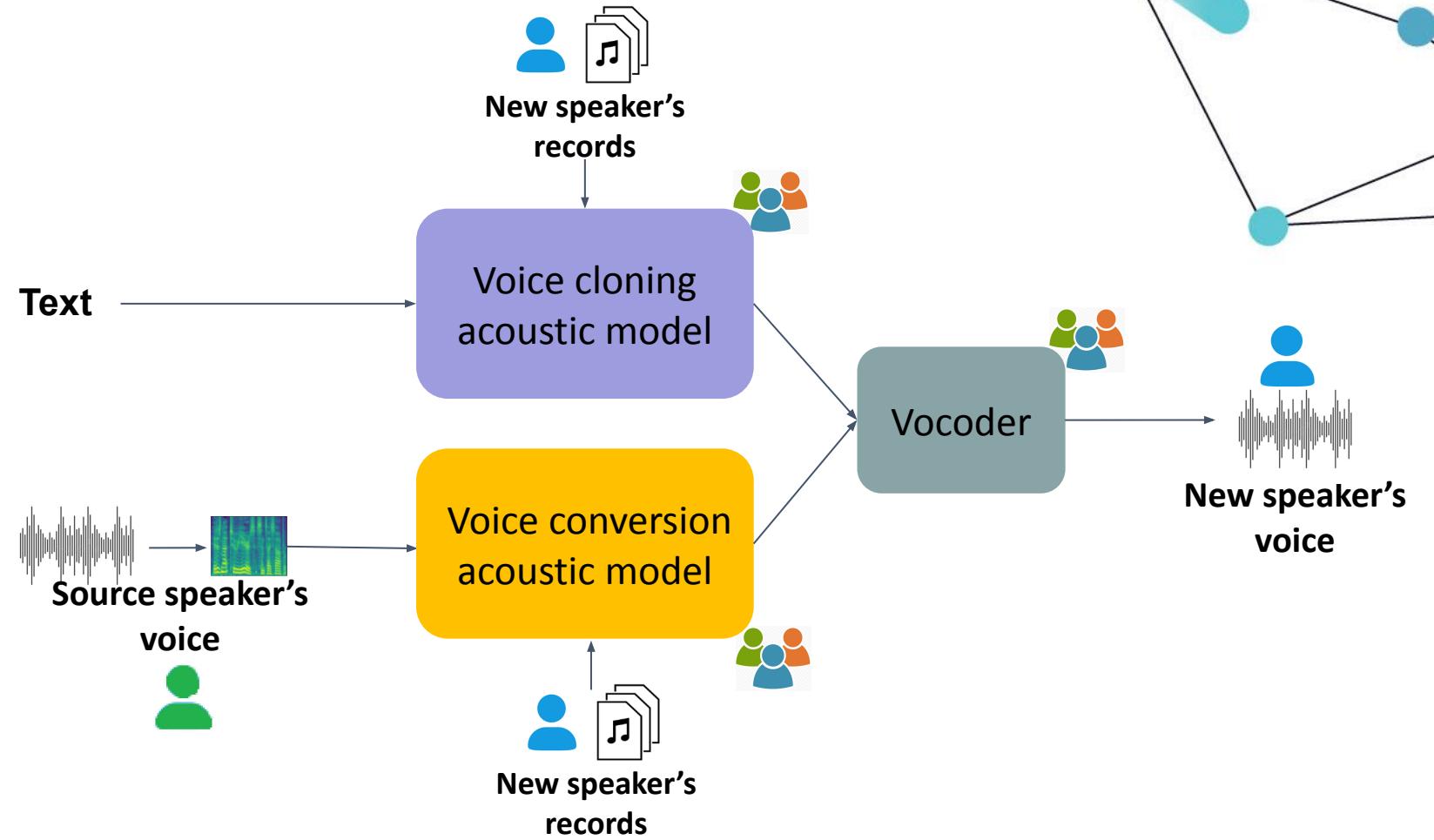


Voice conversion



# Близкие задачи

Voice cloning



Voice conversion

# Метрики качества

# Субъективные метрики качества

Субъективная оценка – **Mean opinion score (MOS)**

Основные критерии:

- naturalness*** – естественность/качество
- speaker similarity*** – похожесть голоса
- \*естественность интонации/соответствие эмоции

Шкала оценки: 1-5

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

# Субъективные метрики качества

Субъективная оценка – **Mean opinion score (MOS)**

Основные критерии:

- naturalness*** – естественность/качество
- speaker similarity*** – похожесть голоса
- \*естественность интонации/соответствие эмоции

Шкала оценки: 1-5

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Процедура оценивания на краудсорсинговых платформах

- протокол ITU-T P.800
- носители языка
- несколько оценщиков на одну запись для более надежных результатов
- добавление ground truth записей из датасета (для верхней границы)
- \*добавление специально зашумленных записей для фильтрации неверных оценок

# Субъективные метрики качества

Субъективная оценка – **Mean opinion score (MOS)**

Основные критерии:

- naturalness*** – естественность/качество
- speaker similarity*** – похожесть голоса
- \*естественность интонации/соответствие эмоции

Шкала оценки: 1-5

Процедура оценивания на краудсорсинговых платформах

- протокол ITU-T P.800
- носители языка
- несколько оценщиков на одну запись для более надежных результатов
- добавление ground truth записей из датасета (для верхней границы) и ресинтезированной вокодером
- \*добавление специально зашумленных записей для фильтрации неверных оценок

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad



Хорошая корреляция с восприятием человека



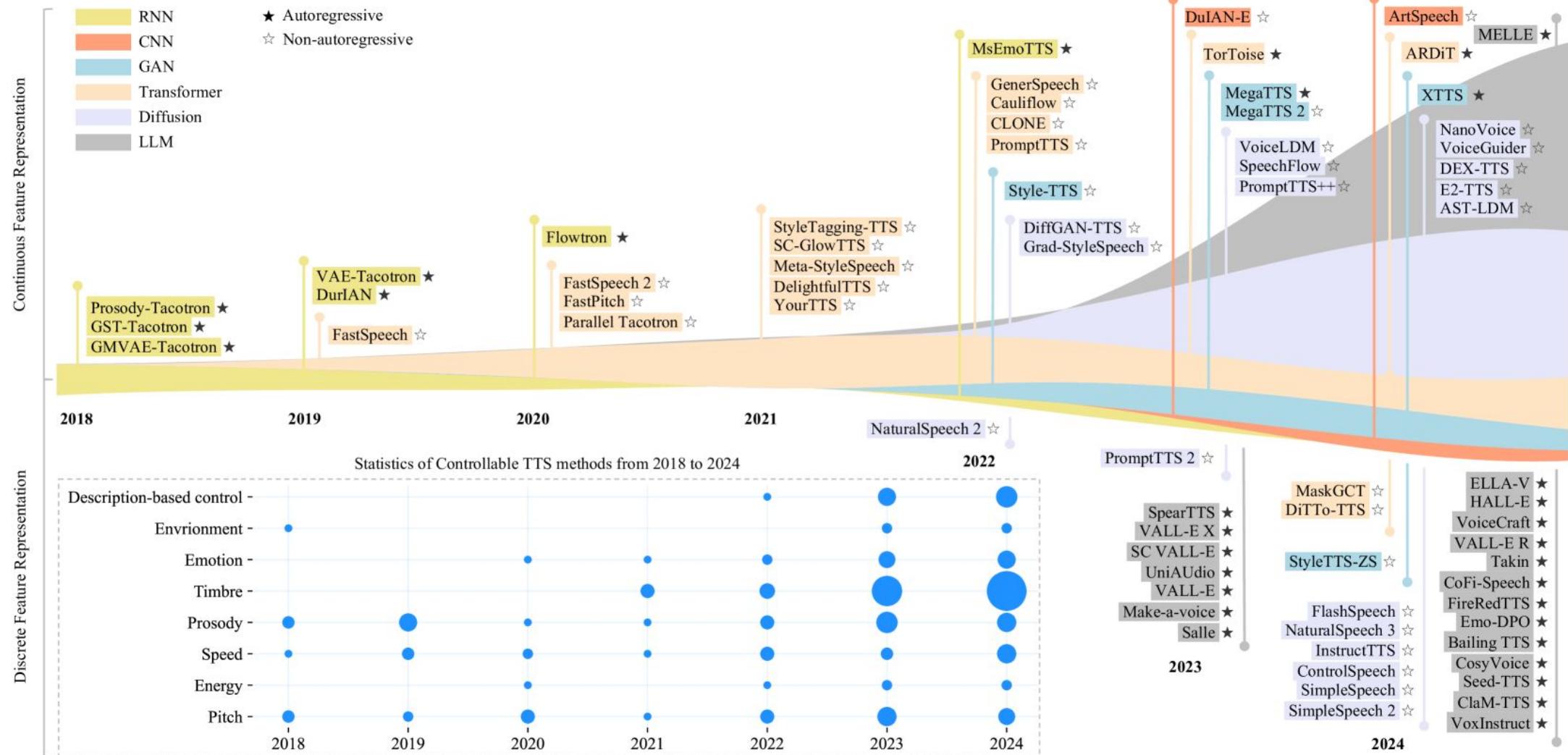
Дорогостоящая процедура  
Требует времени

# Объективные метрики качества

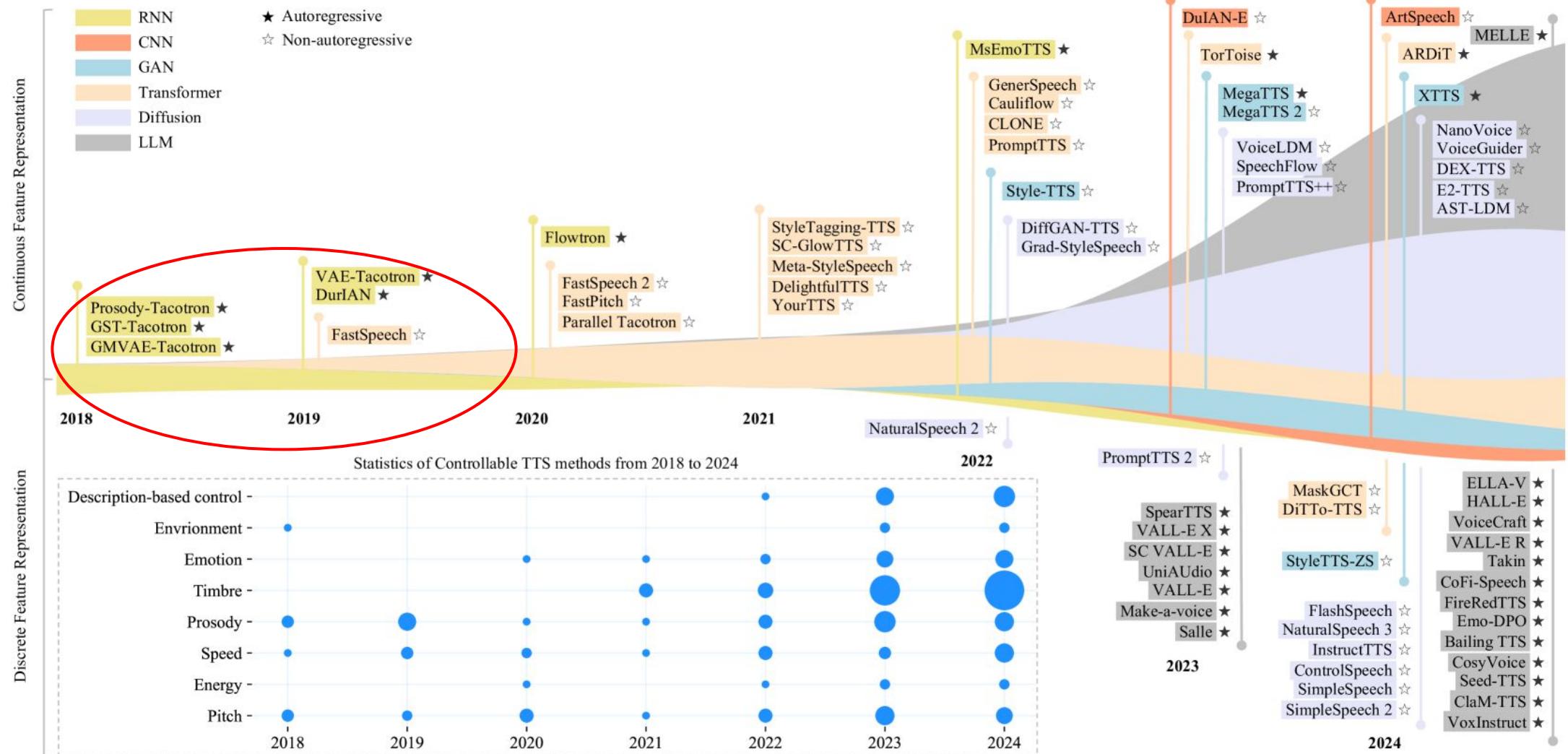
- ❑ Автоматическая оценка MOS
  - например, DNSMOS, UTMOS, NORSEQ
- ❑ Content consistency: word/character error rate (WER/CER)
- ❑ Speaker similarity
  - используется модель, кодирующая голос в speaker embedding (e.g. ERes2Net, wavlm-base-sv)
  - косинусная мера близости между векторами из референсной и сгенерированной записей
- ❑ Оценка эффективности: real time factor (RTF) – время генерации 1 секунды

# Акустические модели

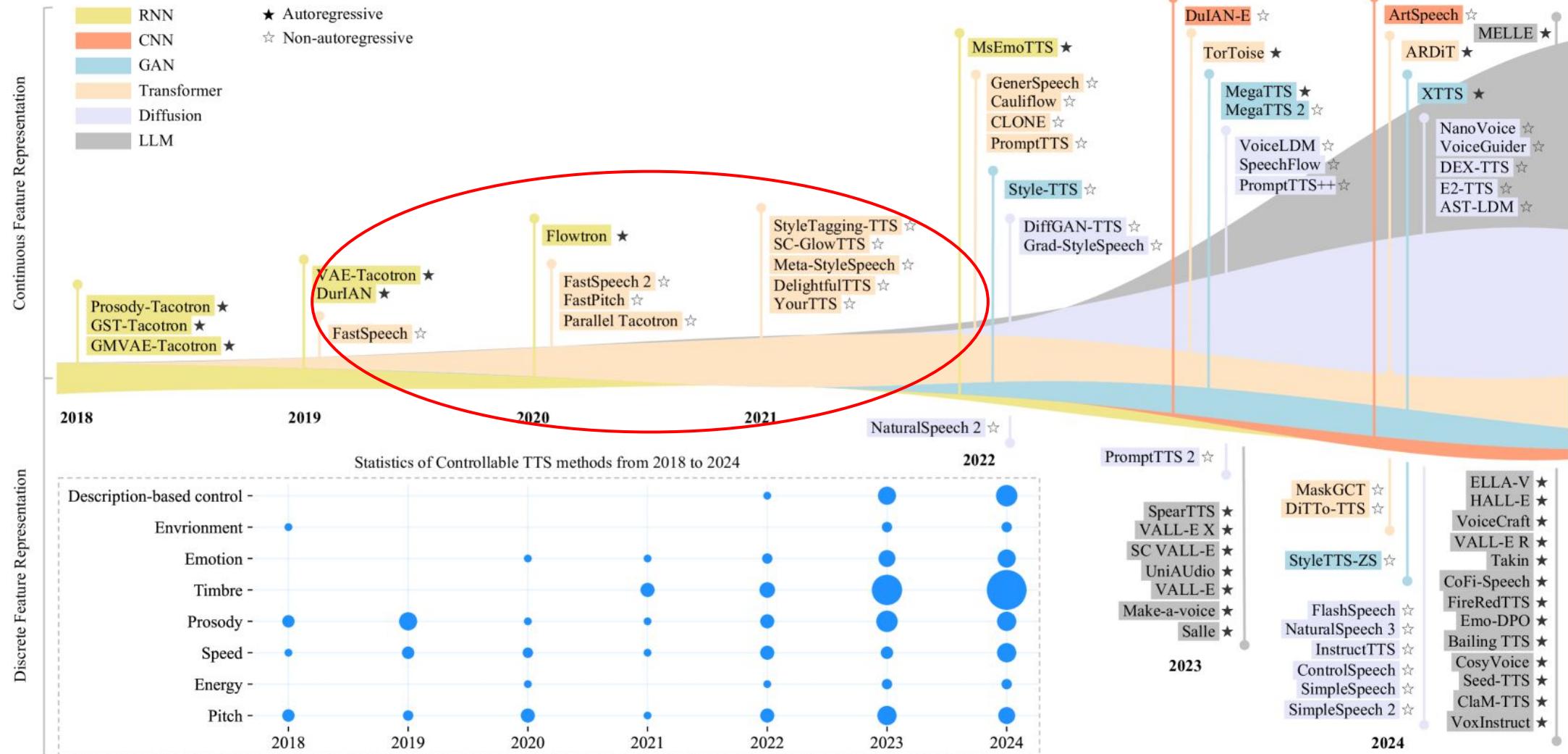
# Эволюция акустических моделей



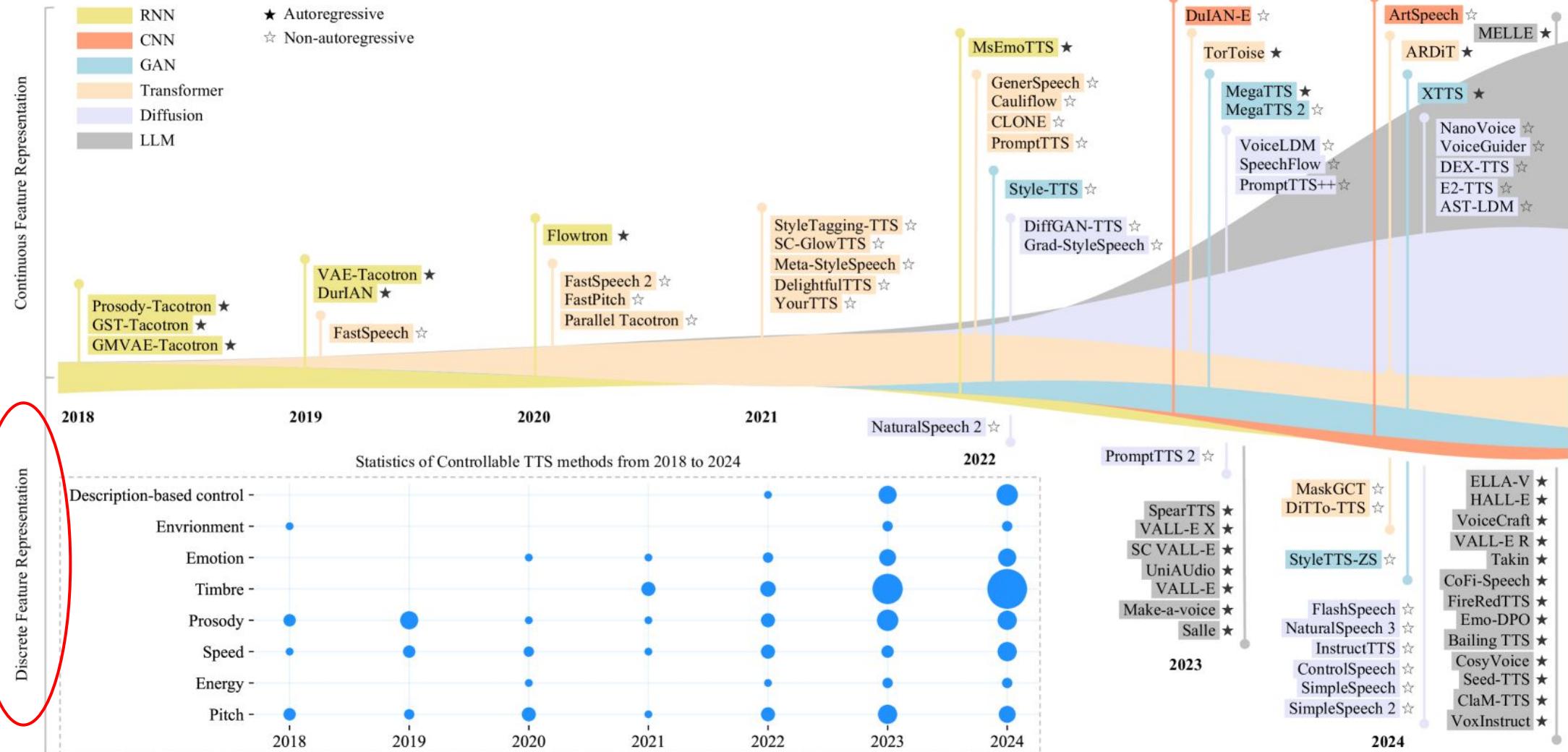
# Эволюция акустических моделей



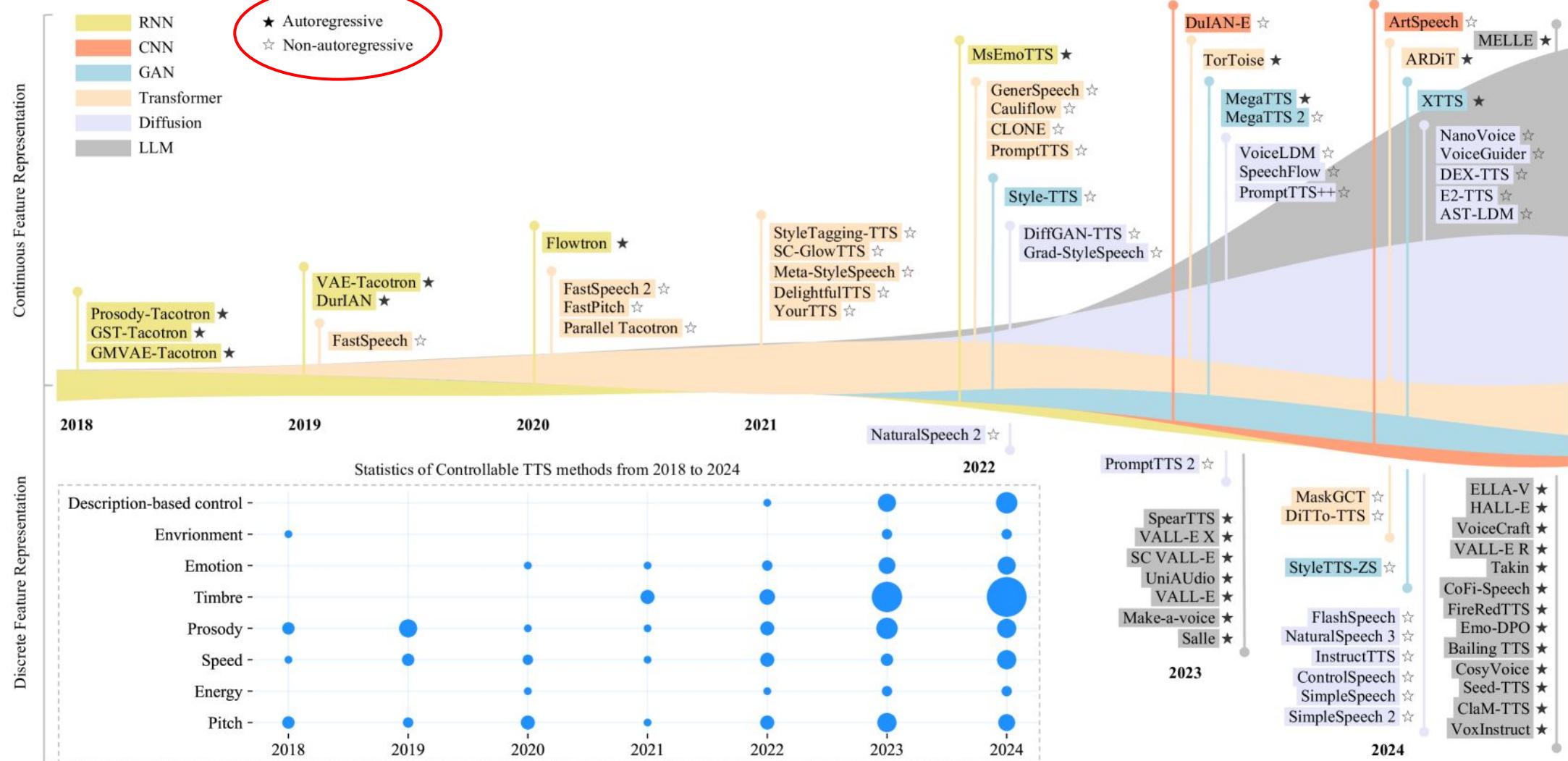
# Эволюция акустических моделей



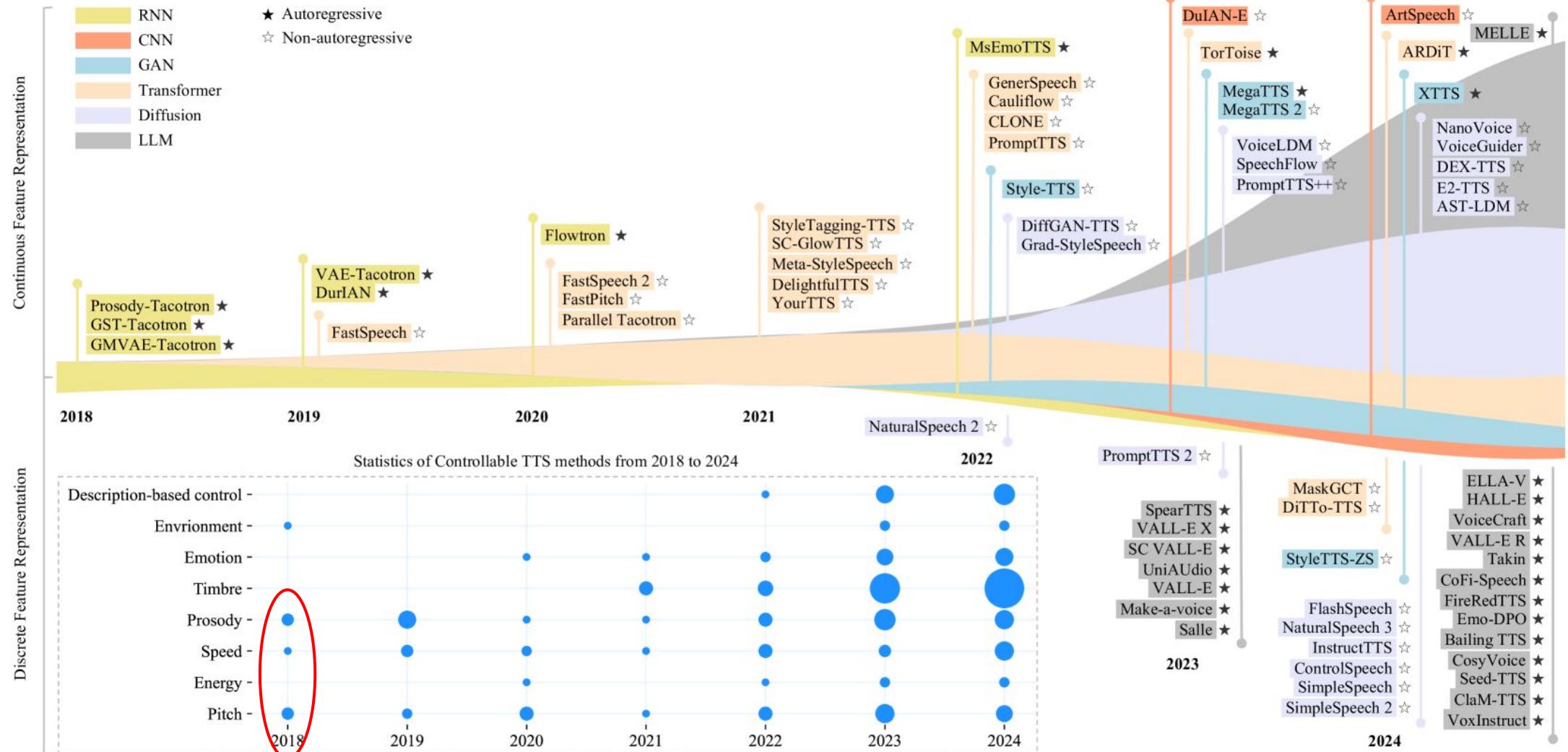
# Эволюция акустических моделей



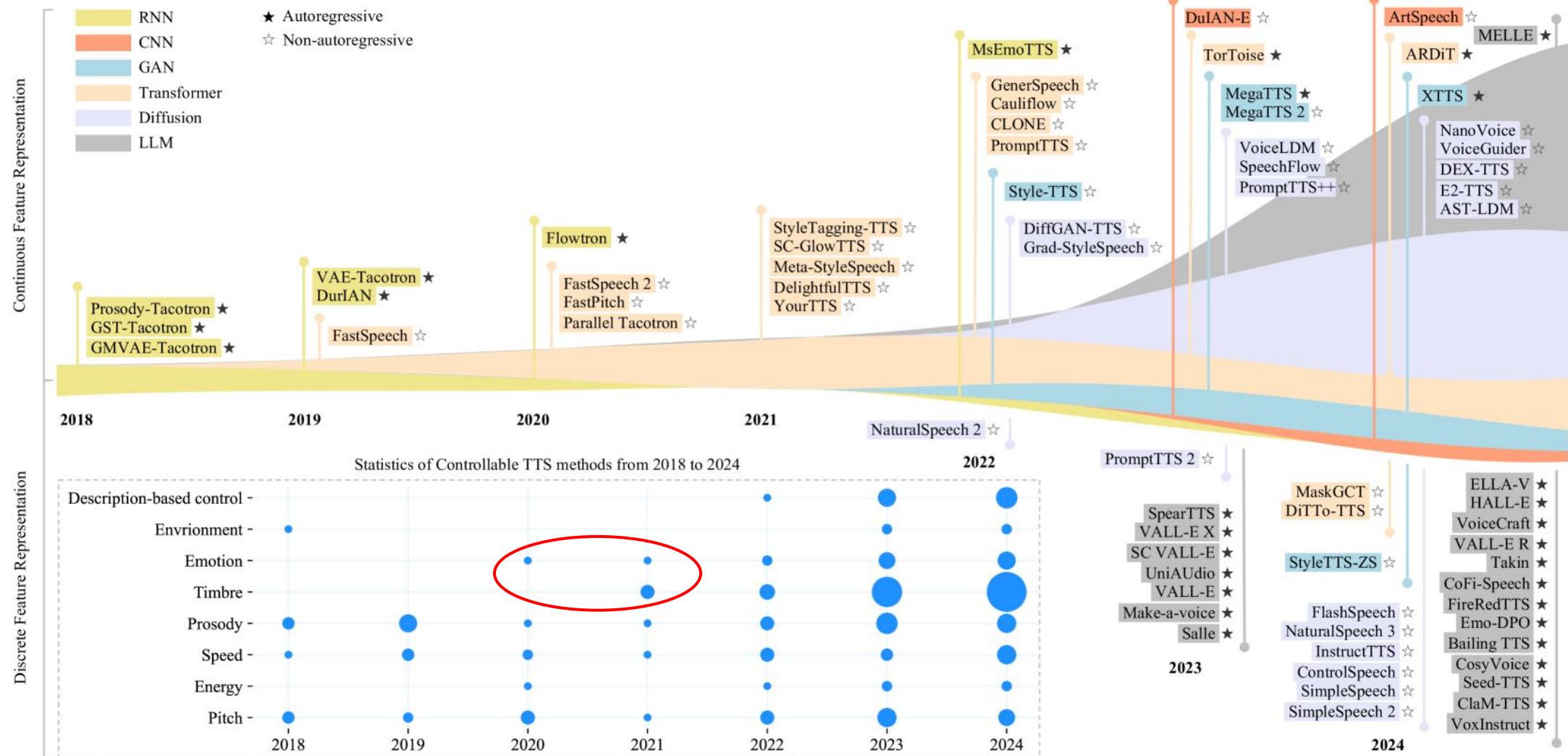
# Эволюция акустических моделей



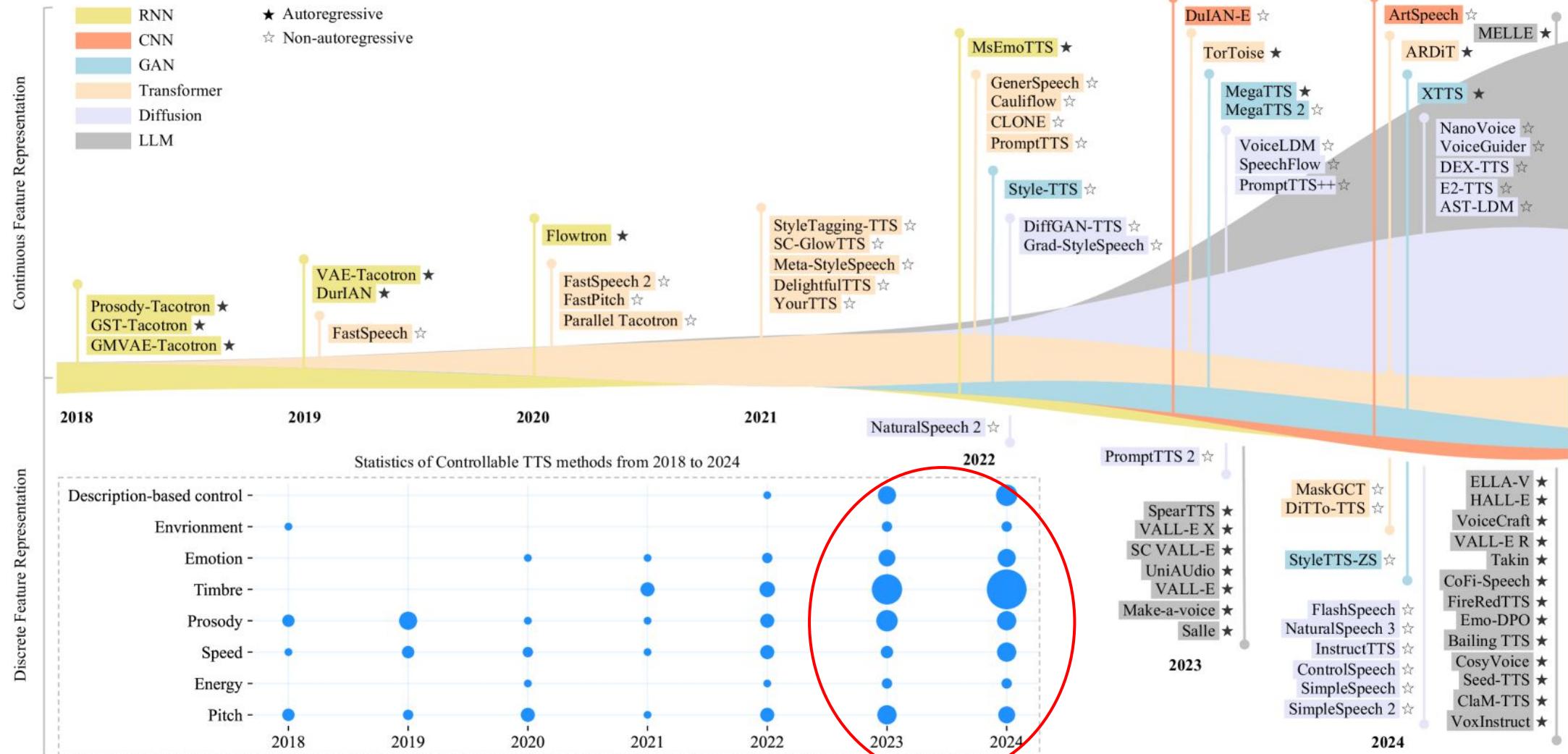
# Эволюция акустических моделей



# Эволюция акустических моделей

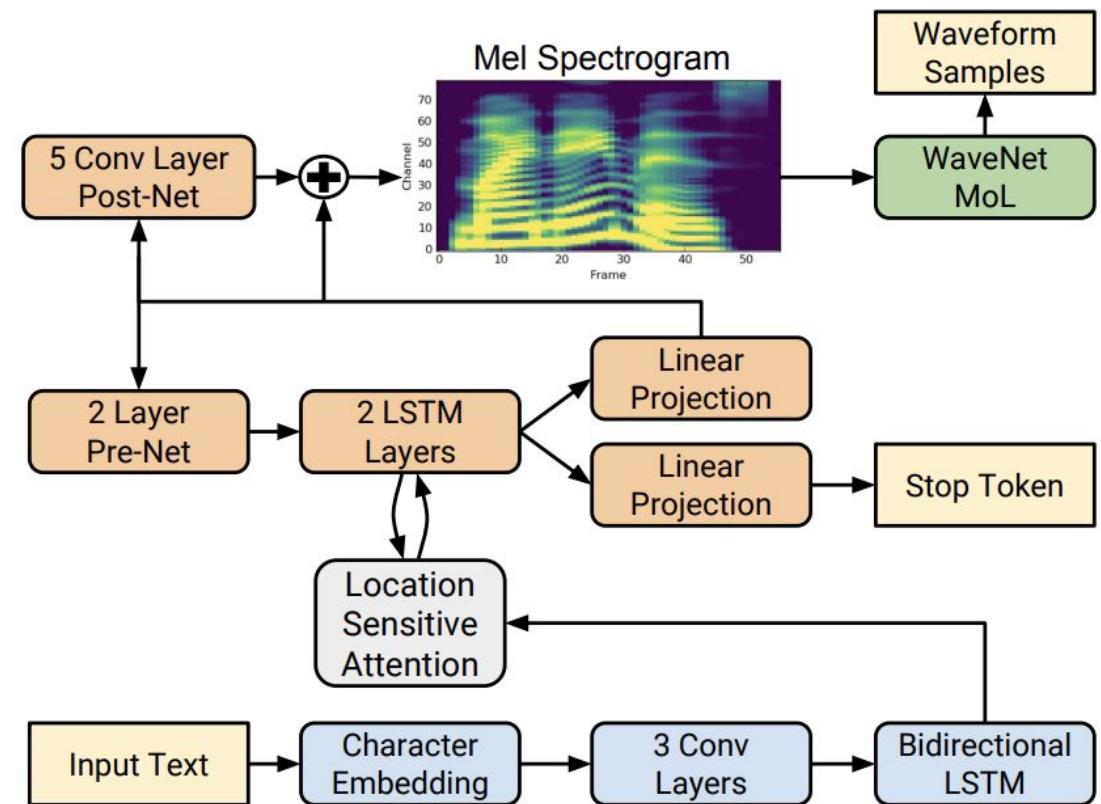


# Эволюция акустических моделей



# Tacotron2 (2018)

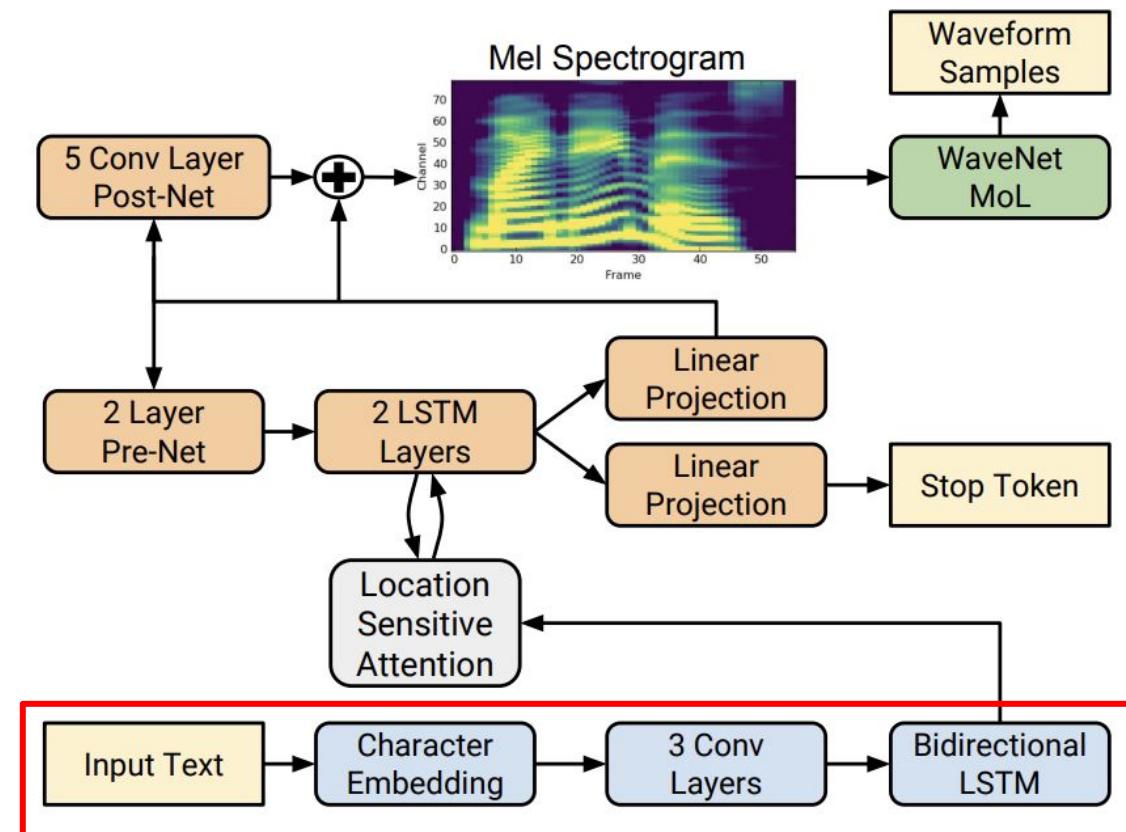
- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель



[Статья](#), [github](#), [github](#)

# Tacotron2 (2018)

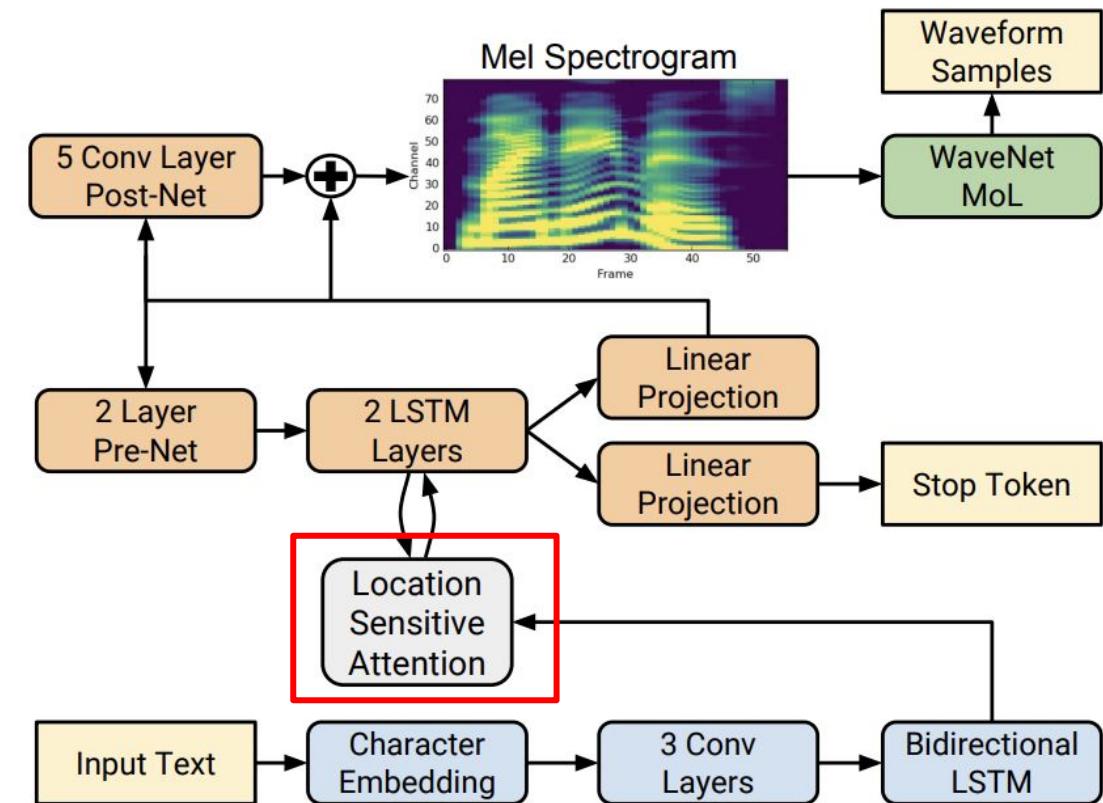
- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста



[Статья](#), [github](#), [github](#)

# Tacotron2 (2018)

- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста
  2. **Attention** – выравнивание между текстом и фреймами



[Статья](#), [github](#), [github](#)

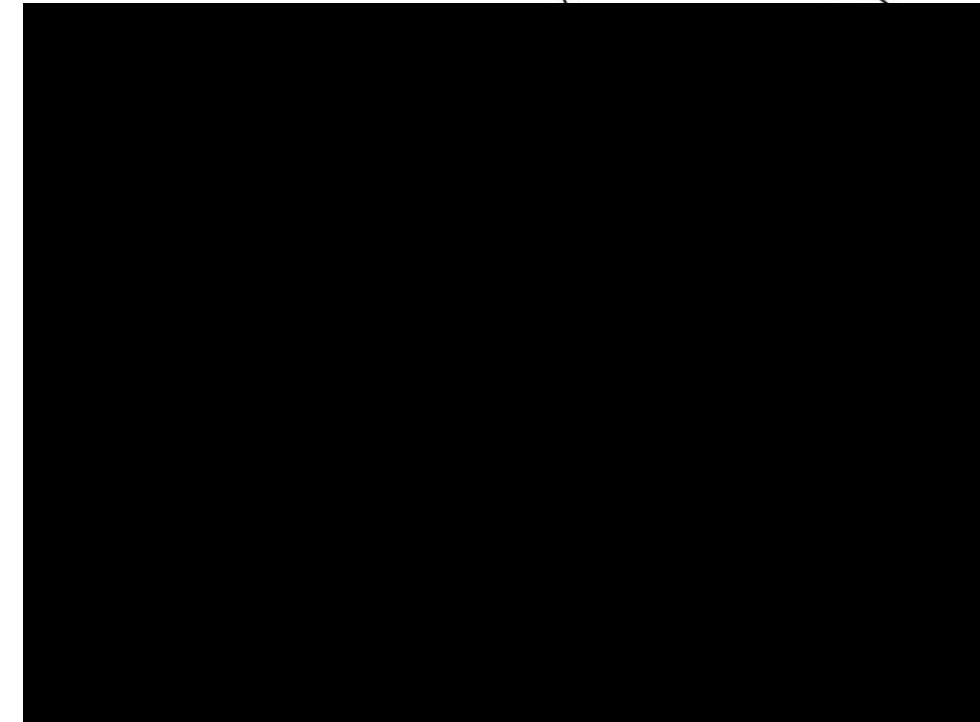
# Tacotron2 (2018)

- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста
  2. **Attention** – выравнивание между текстом и фреймами
    - $h$  – encoder hidden states,  $T_x$  – длина текста
    - $s$  – decoder hidden states
    - $c$  – context vector

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \text{softmax}(\text{score}(s_i, h_j))$$

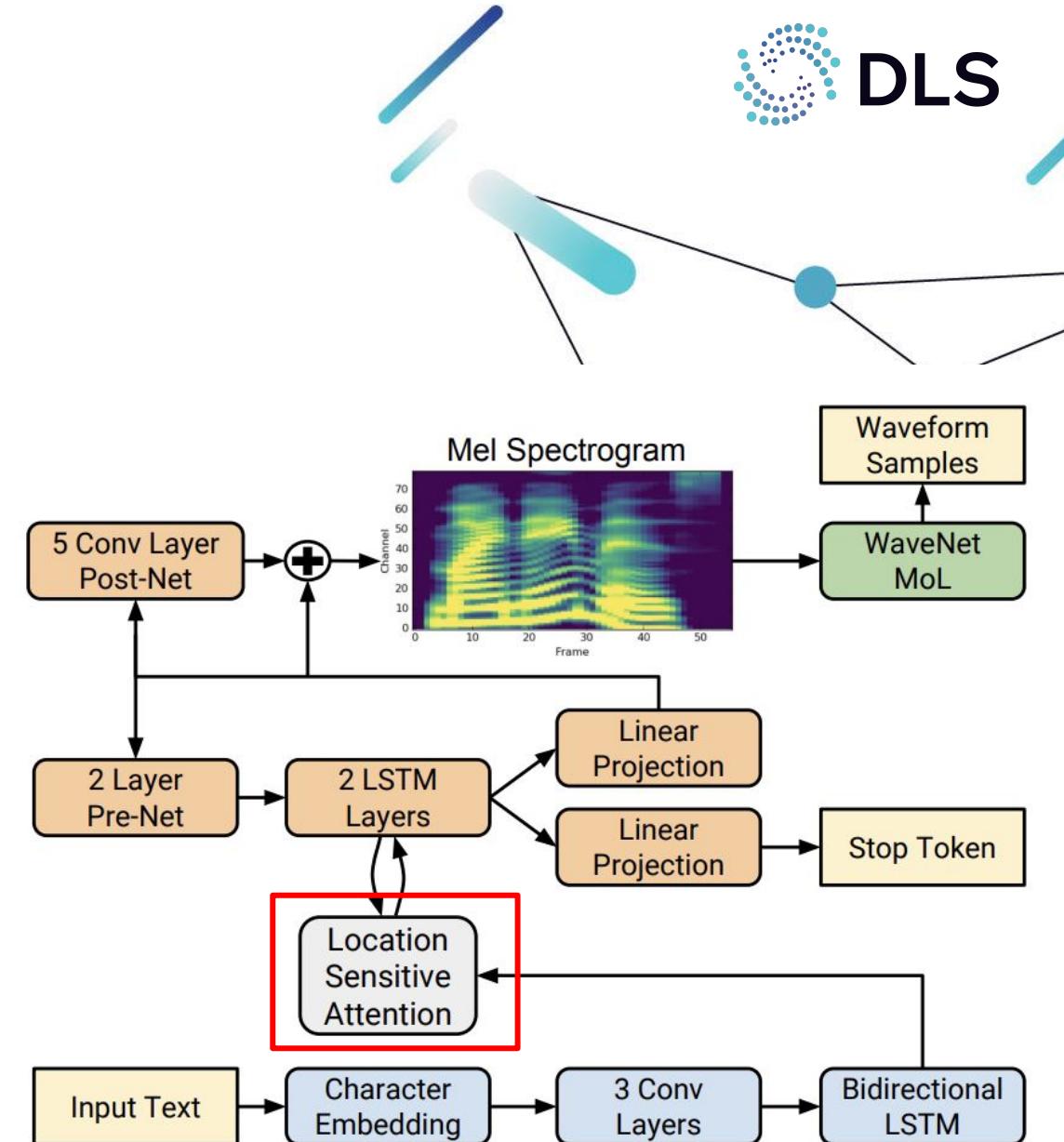
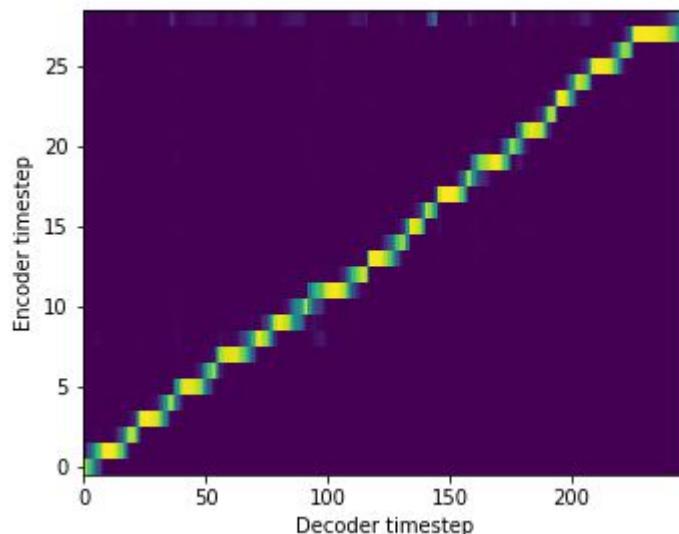
$$\text{score}(s_i, h_j) = \begin{cases} s_i^T h_j, & \text{dot-product} \\ s_i W h_j, & \text{multiplicative} \\ W_3^T \tanh(W_1 s_i + W_2 h_j) & \text{additive} \end{cases}$$



видео с [сайта](#)

# Tacotron2 (2018)

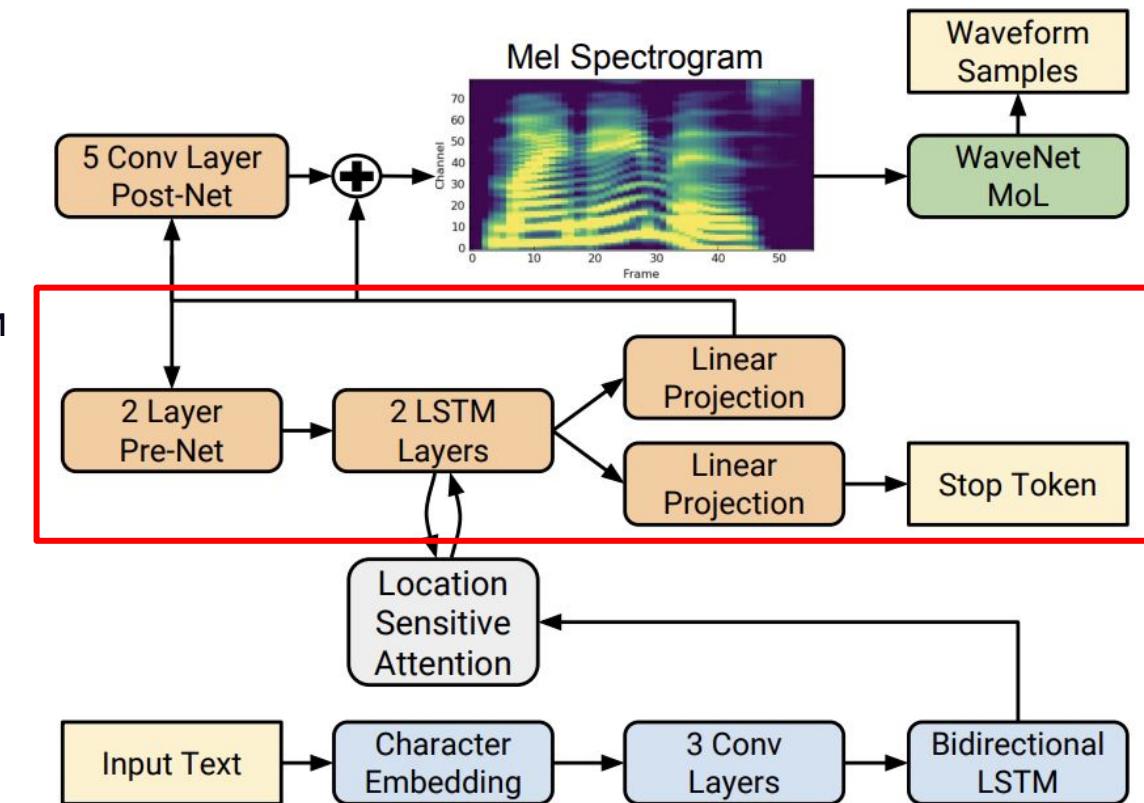
- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста
  2. **Attention** – выравнивание между текстом и фреймами  
*Location sensitive attention*



[Статья](#), [github](#), [github](#)

# Tacotron2 (2018)

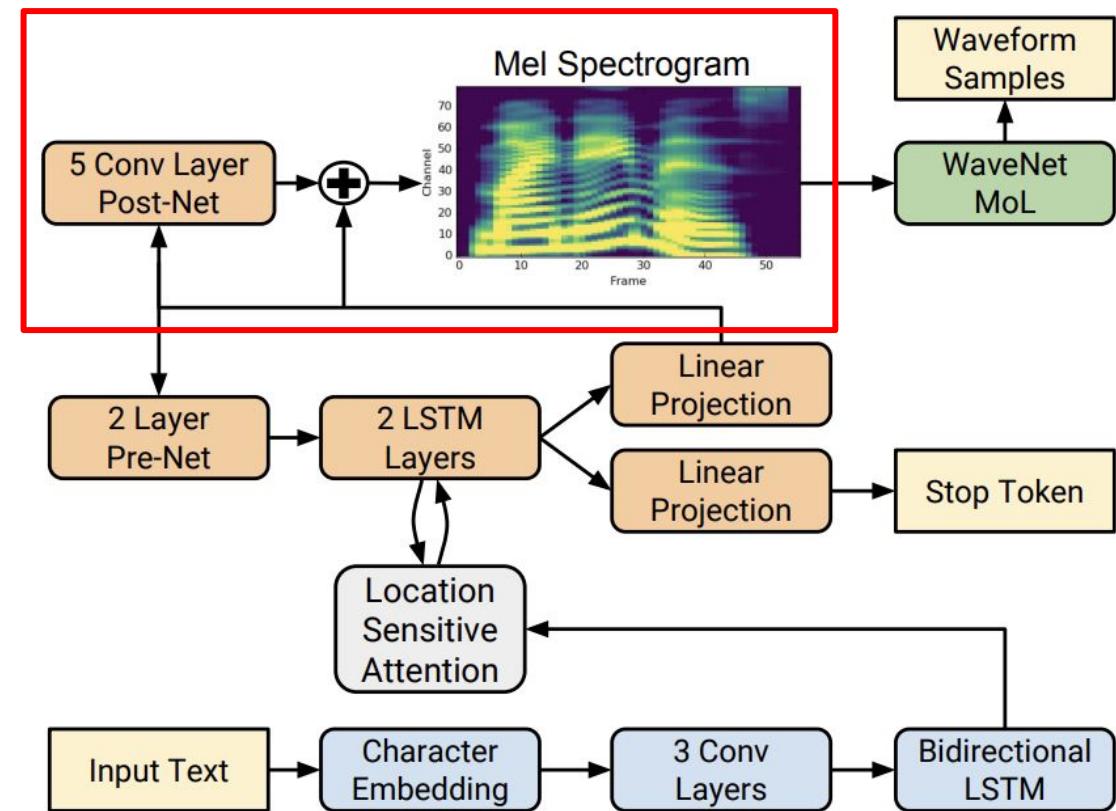
- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста
  2. **Attention** – выравнивание между текстом и фреймами
  3. **Decoder** (LSTM) – авторегрессионное предсказание  
вход – предыдущие предсказания + context vector  
выход – фрейм спектrogramмы + стоп токена



[Статья](#), [github](#), [github](#)

# Tacotron2 (2018)

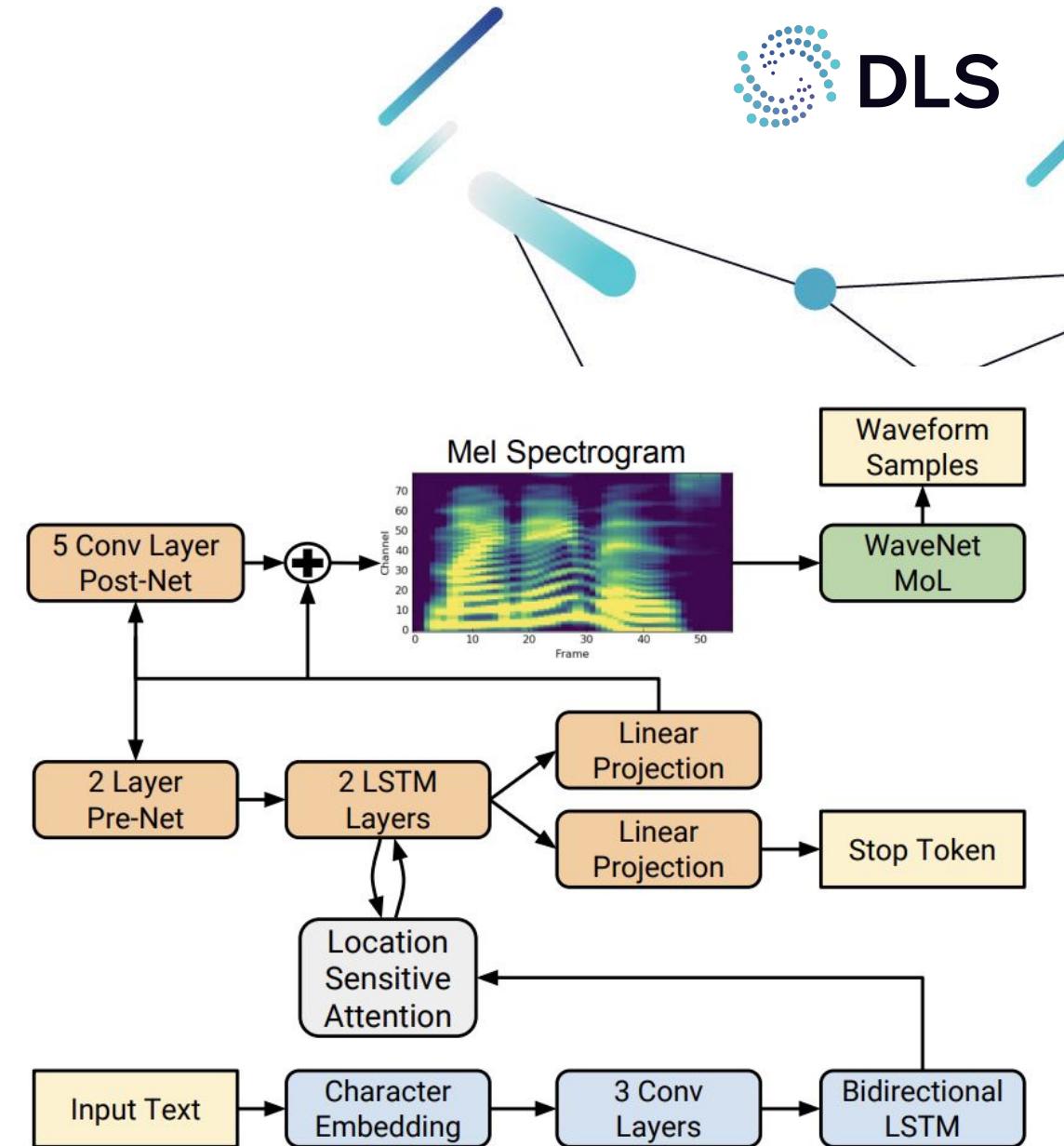
- ❑ Семейство моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста
  2. **Attention** – выравнивание между текстом и фреймами
  3. **Decoder** (LSTM) – авторегрессионное предсказание  
вход – предыдущие предсказания + attention vector  
выход – фрейм спектrogramмы + стоп токена
  4. **Postnet** (CNN + residual connection) – улучшение  
качества



[Статья](#), [github](#), [github](#)

# Tacotron2 (2018)

- ❑ Семейство RNN-based моделей, первая версия в 2017
- ❑ Вход – буквы, выход – мел-спектrogramма
- ❑ Авторегрессионная модель
- ❑ 4 основных блока:
  1. **Encoder** (CNN+bi-LSTM) – обработка текста
  2. **Attention** – выравнивание между текстом и фреймами
  3. **Decoder** (LSTM) – авторегрессионное предсказание
   
вход – предыдущие предсказания + attention vector
   
выход – фрейм спектrogramмы + стоп токена
  4. **Postnet** (CNN + residual connection) – улучшение
   
качества
- ❑ Teacher-forcing
- ❑ Loss: MSE между GT спектrogramмой и полученной до и
   
после Postnet + кросс-энтропия на предсказания стоп токена
- ❑ Вокодер WaveNet (CNN, autoregressive)



[Статья](#), [github](#), [github](#)

# Tacotron2 (2018)

- ❑ Ресурсы: 1 GPU
- ❑ Данные: 24.6 часа, один женский голос
- ❑ Размер: 28.2 М
- ❑ RTF: ~0.25
- ❑ Naturalness:

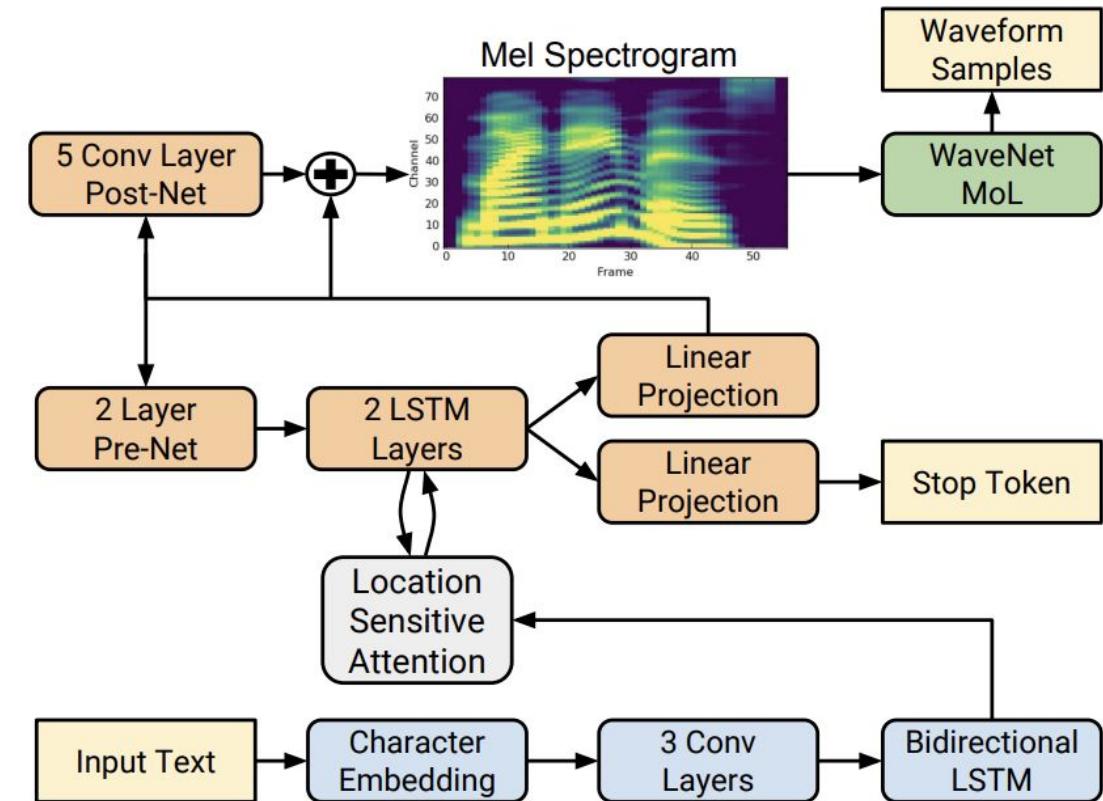
System	MOS
Parametric	$3.492 \pm 0.096$
Tacotron (Griffin-Lim)	$4.001 \pm 0.087$
Concatenative	$4.166 \pm 0.091$
WaveNet (Linguistic)	$4.341 \pm 0.051$
Ground truth	$4.582 \pm 0.053$
Tacotron 2 (this paper)	<b><math>4.526 \pm 0.066</math></b>

Tacotron2



"Don't desert me here in the desert!"

[Demo](#)



[Статья](#), [github](#), [github](#)

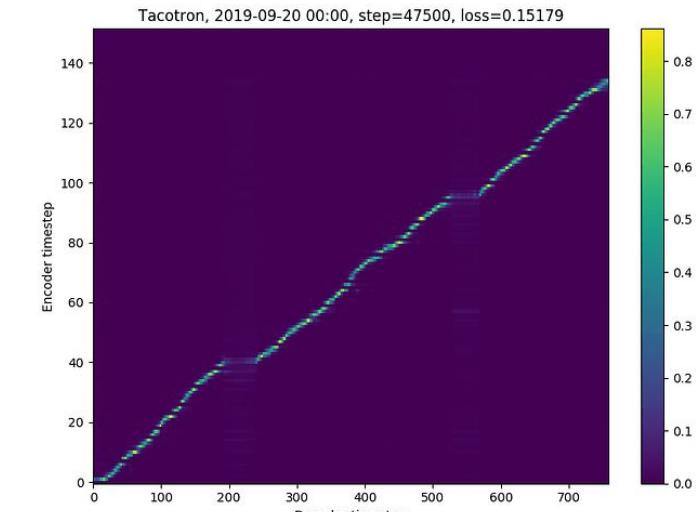
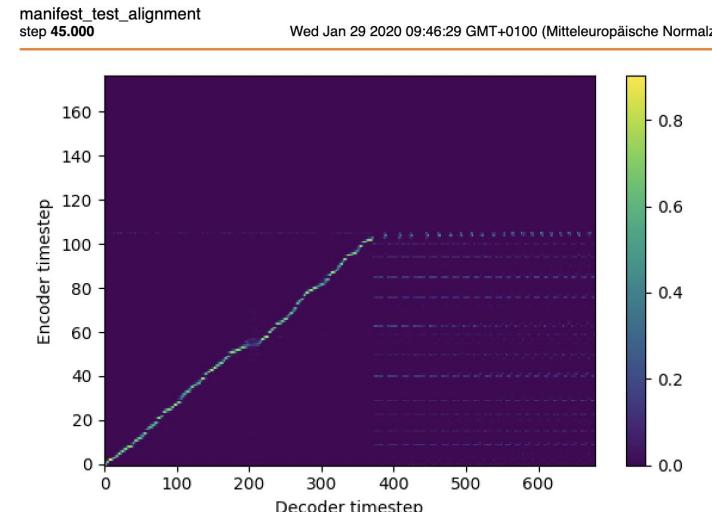
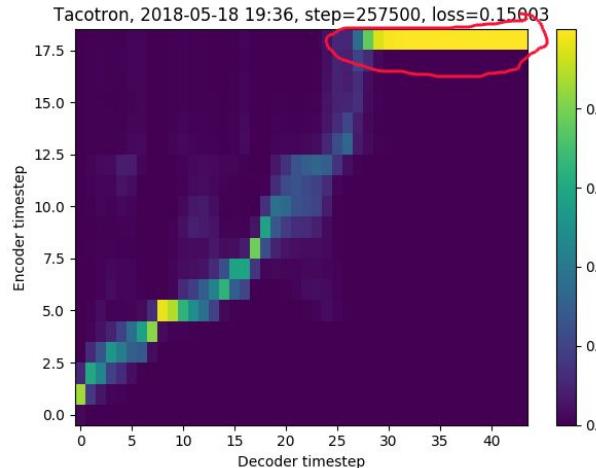
# Tacotron2 (2018)



- MOS лучше параметрической и конкатенативной систем
- Хорошо звучащие записи для одного голоса



- Нестабильное обучение из-за attention
- Проблемы с четкостью произношения, паузами, концом синтеза



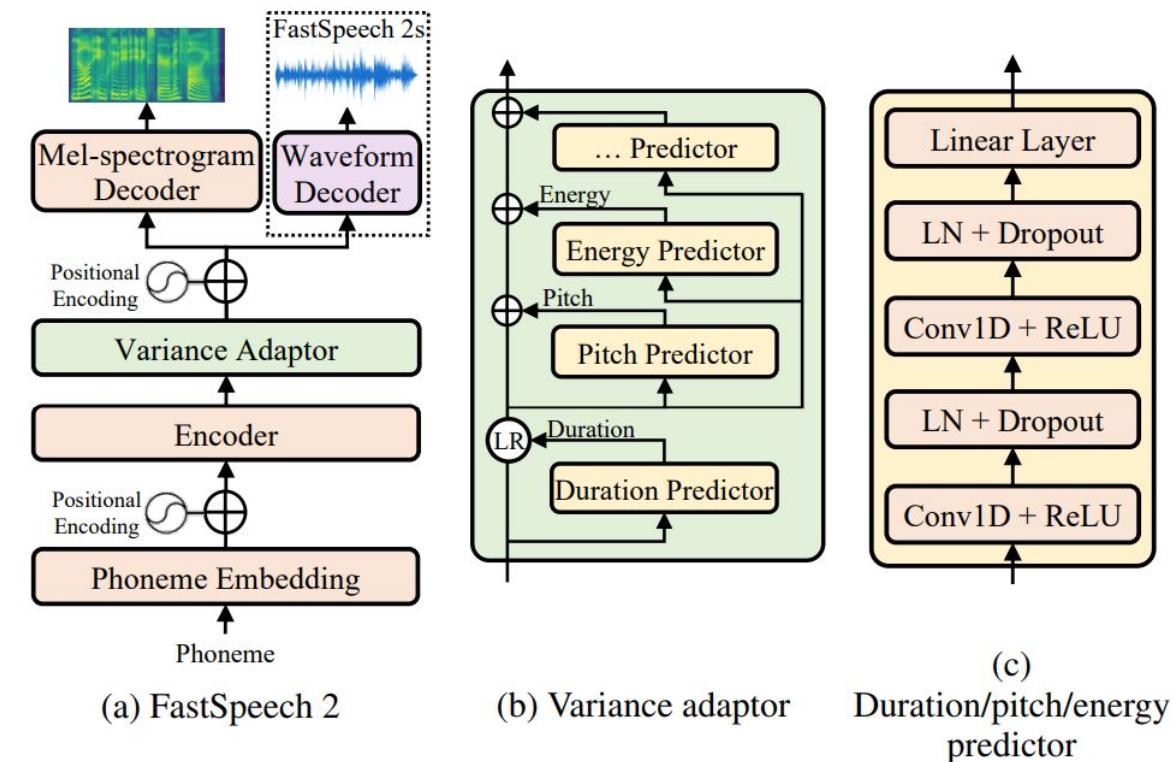
# Улучшения модели Tacotron2

- ❑ Замена LSA attention
  - на другие типы: e.g. stepwise monotonic attention ([paper](#))
  - на duration predictor: Non-attentive Tacotron ([paper](#))
- ❑ Контроль просодией
  - Prosody-Tacotron ([paper](#))
  - Tacotron-GST ([paper](#))
  - GMVAE-Tacotron ([paper](#))
- ❑ Генерация для нескольких языков ([paper](#))

До сих пор используется в продуктowych решениях, хоть и со значительными модификациями

# FastSpeech2 (2020)

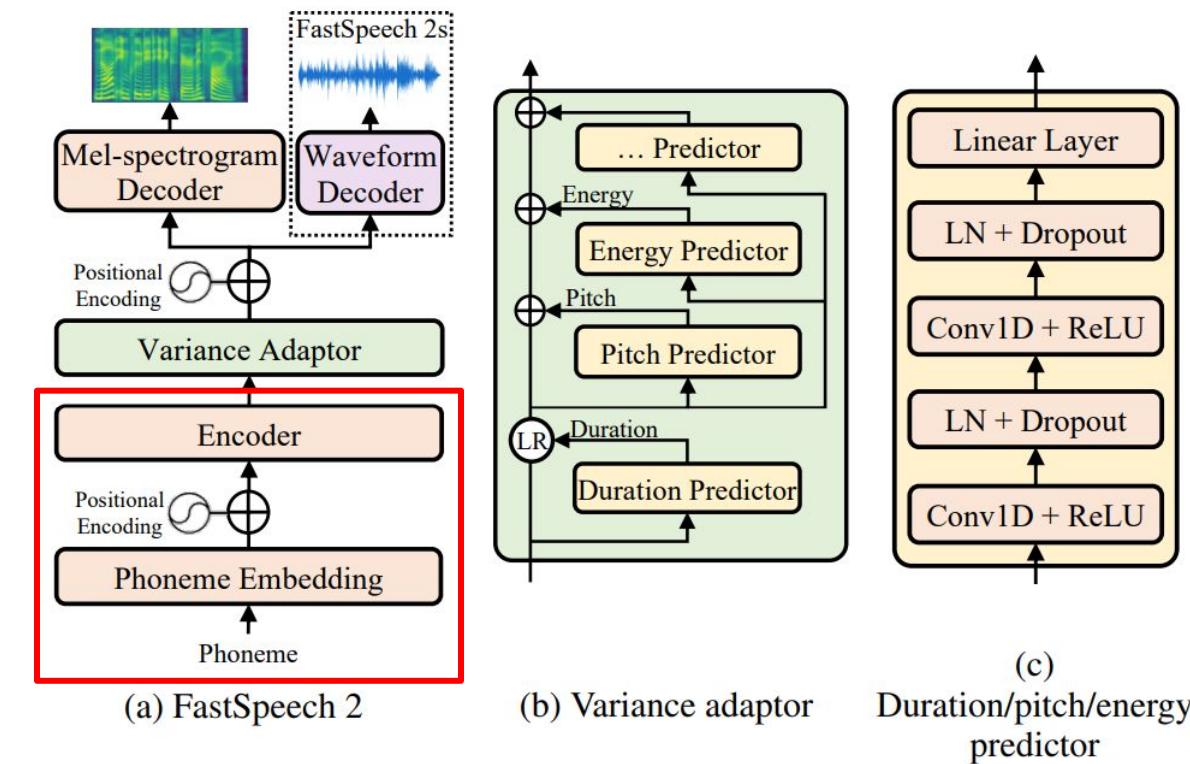
- ❑ Семейство Transformer-based моделей
- ❑ Неавторегрессионная
- ❑ Вход – фонемы, выход – мел-спектrogramма



[статья](#), [github](#), [huggingface](#)

# FastSpeech2 (2020)

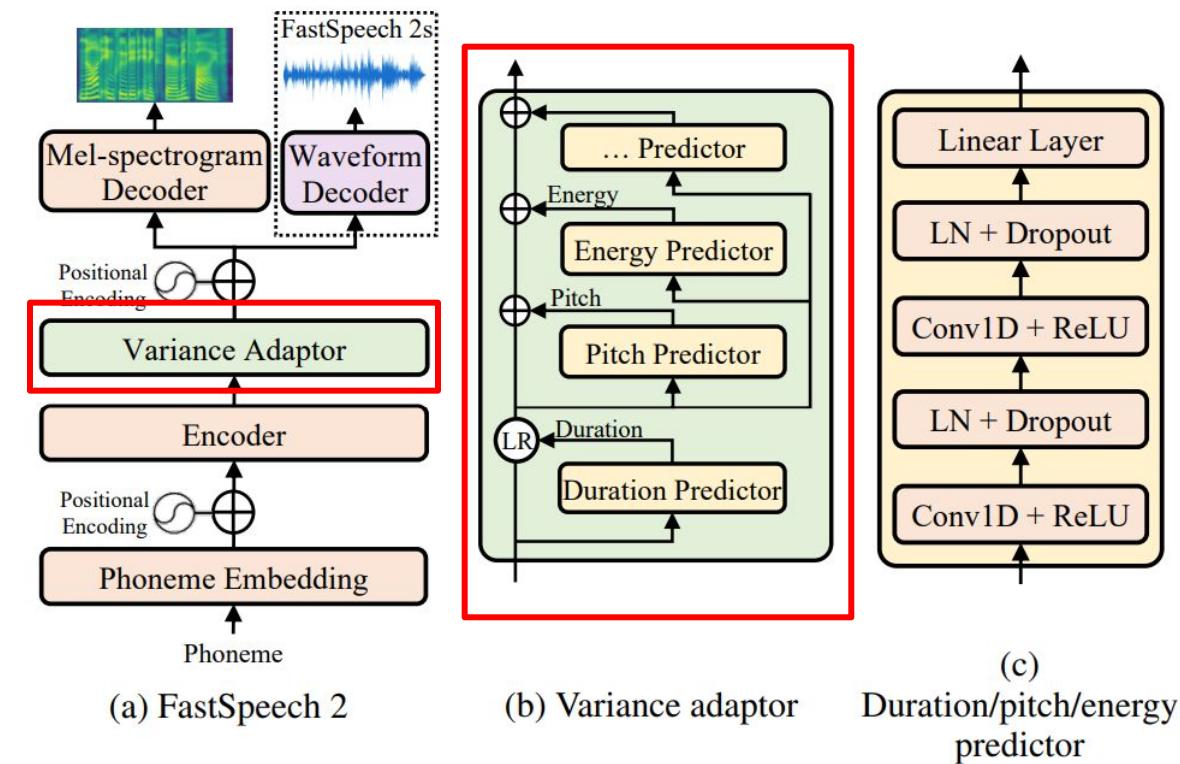
- ❑ Семейство Transformer-based моделей
- ❑ Неавторегрессионная
- ❑ Вход – фонемы, выход – мел-спектrogramма
- ❑ 3 блока:
  1. **Encoder** обрабатывает фонемы



[статья](#), [github](#), [huggingface](#)

# FastSpeech2 (2020)

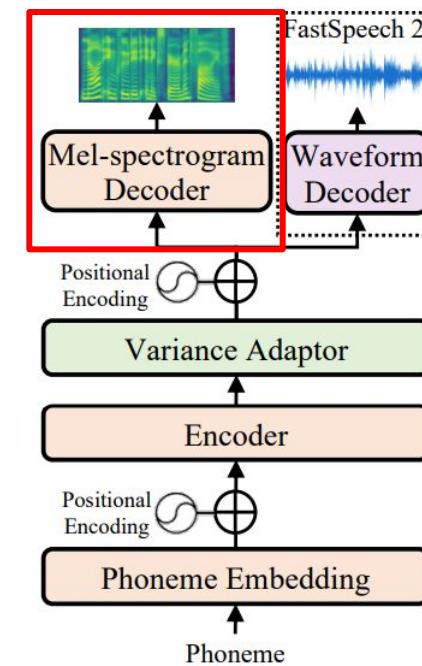
- ❑ Семейство Transformer-based моделей
- ❑ Неавторегрессионная
- ❑ Вход – фонемы, выход – мел-спектrogramма
- ❑ 3 блока:
  1. **Encoder** обрабатывает фонемы
  2. **Variance adaptors**
    - *duration predictor* – предсказывает для каждой фонемы ее длительность
    - *pitch/energy predictor* - улучшение просодии



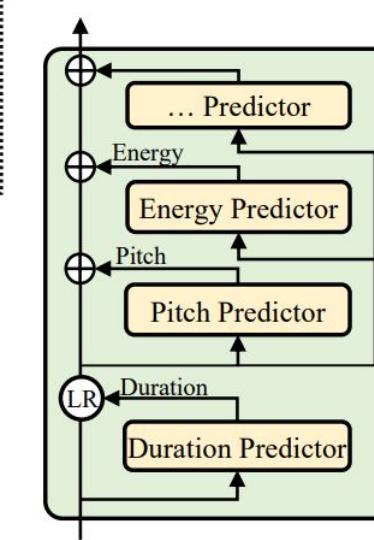
[статья](#), [github](#), [huggingface](#)

# FastSpeech2 (2020)

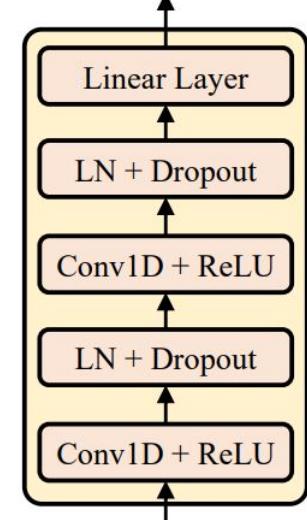
- ❑ Семейство Transformer-based моделей
- ❑ Неавторегрессионная
- ❑ Вход – фонемы, выход – мел-спектrogramма
- ❑ 3 блока:
  1. **Encoder** обрабатывает фонемы
  2. **Variance adaptors**
    - *duration predictor* – предсказывает для каждой фонемы ее длительность
    - *pitch/energy predictor* - улучшение просодии
  3. **Decoder** генерирует спектrogramму за один проход



(a) FastSpeech 2



(b) Variance adaptor

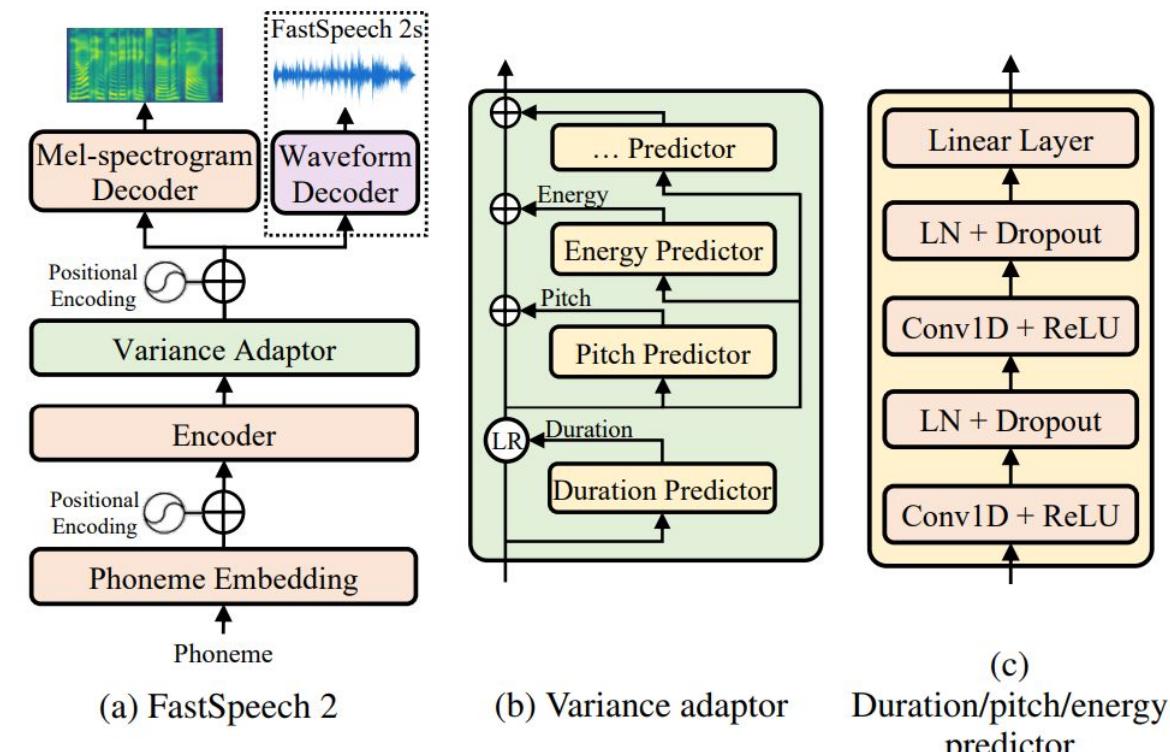


(c)  
Duration/pitch/energy  
predictor

[статья](#), [github](#), [huggingface](#)

# FastSpeech2 (2020)

- Семейство Transformer-based моделей
- Неавторегрессионная
- Вход – фонемы, выход – мел-спектrogramма
- 3 блока:
  1. **Encoder** обрабатывает фонемы
  2. **Variance adaptors**
    - *duration predictor* – предсказывает для каждой фонемы ее длительность
    - *pitch/energy predictor* - улучшение просодии
  3. **Decoder** генерирует спектrogramму за один проход
- Teacher forcing для variance adaptors
- Loss:
  - MAE на спектrogramмы
  - MSE на длительности, pitch, energy (нужны gt значения)
- Вокодер Parallel WaveGAN



[статья](#), [github](#), [huggingface](#)

# FastSpeech2 (2020)

- Ресурсы: 1 GPU
- Данные: 24 часа, один женский голос (LJSpeech)
- Размер: 24M
- RTE: ~0.02 (with vocoder)
- Naturalness:

Method	MOS
<i>GT</i>	$4.30 \pm 0.07$
<i>GT (Mel + PWG)</i>	$3.92 \pm 0.08$
<i>Tacotron 2 (Shen et al., 2018) (Mel + PWG)</i>	$3.70 \pm 0.08$
<i>Transformer TTS (Li et al., 2019) (Mel + PWG)</i>	$3.72 \pm 0.07$
<i>FastSpeech (Ren et al., 2019) (Mel + PWG)</i>	$3.68 \pm 0.09$
<i>FastSpeech 2 (Mel + PWG)</i>	$3.83 \pm 0.08$
<i>FastSpeech 2s</i>	$3.71 \pm 0.09$

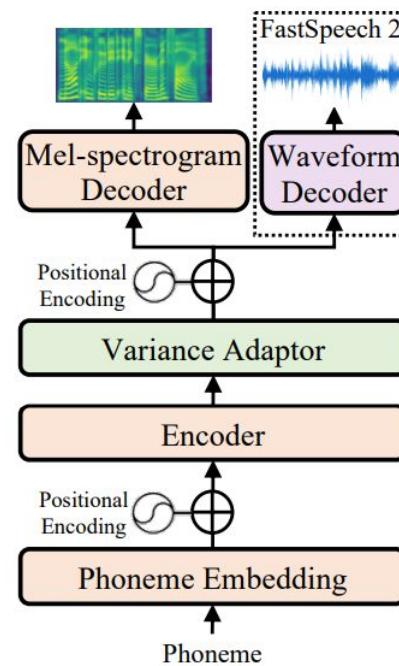
GT



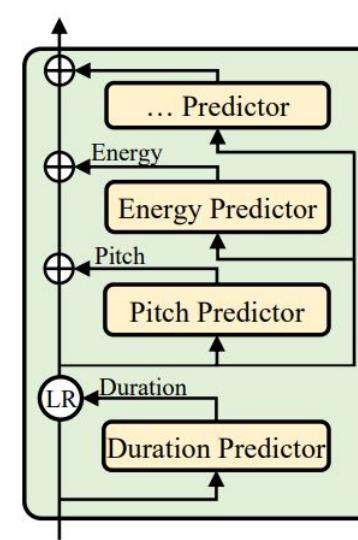
GT (Mel+PWG)



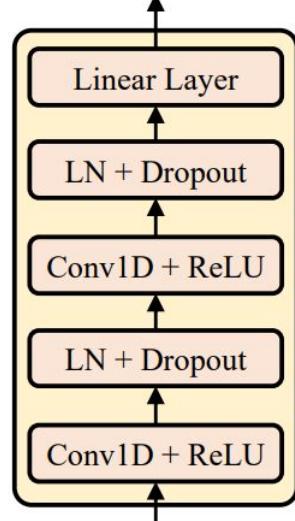
FastSpeech



(a) FastSpeech 2



(b) Variance adaptor



(c) Duration/pitch/energy predictor

[статья](#), [github](#), [huggingface](#)
[Demo](#)

# FastSpeech2 (2020)



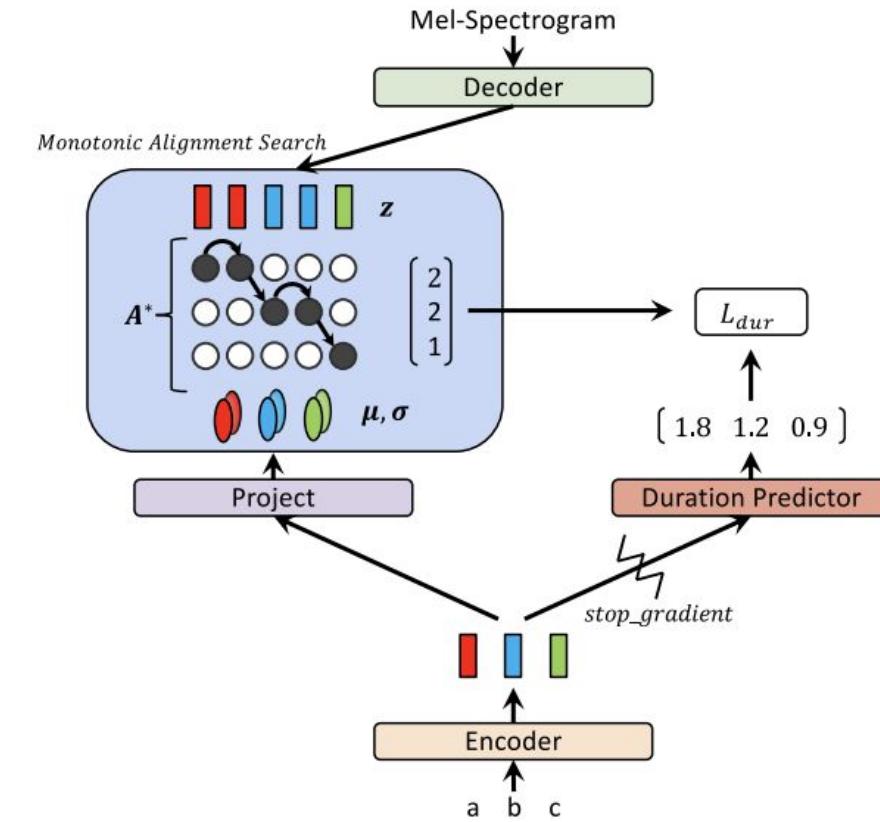
- Быстрая (неавторегрессионная)
- Разнообразная речь, можно в некоторой степени манипулировать питчом и длительностями
- Стабильная



- Для обучения необходимо знать  $gt$  значения длительностей

# Glow-TTS (2020)

- ❑ Flow-based модель
- ❑ Неавторегрессионная
- ❑ 3 основных блока:
  1. **Encoder**: Transformer, фонемы  
Выход – статистики  $\mu, \sigma$  для каждой фонемы

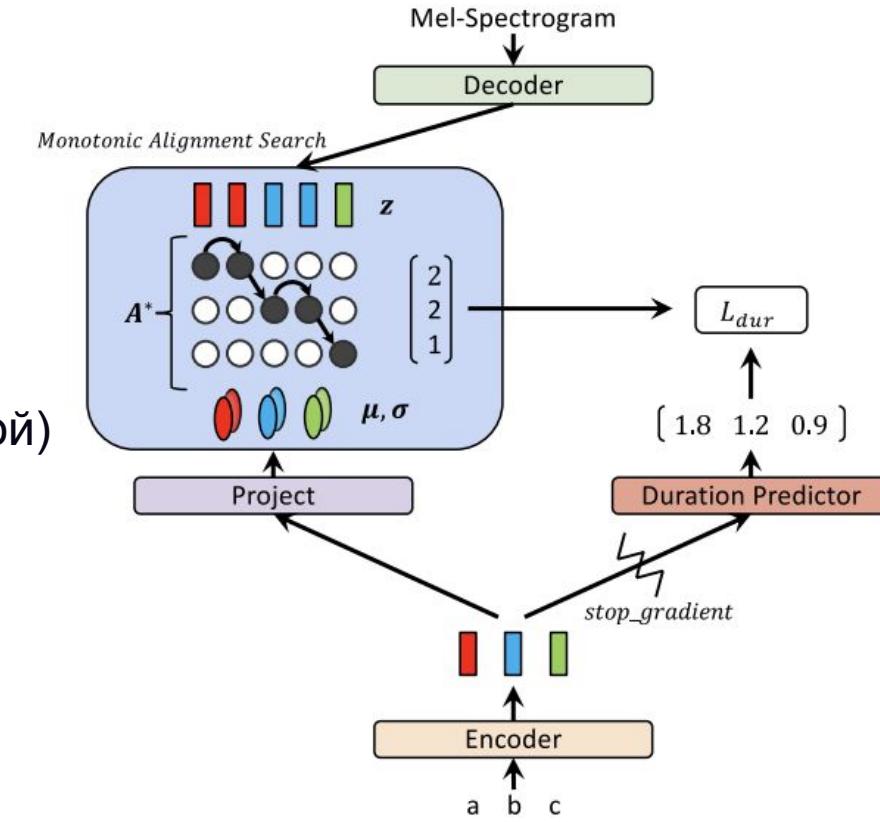
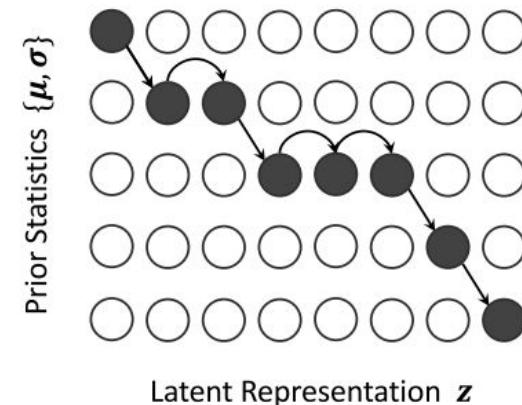


(a) An abstract diagram of the training procedure.

[Статья](#), [github](#)

# Glow-TTS (2020)

- ❑ Flow-based модель
- ❑ 3 основных блока:
  1. Encoder: Transformer, фонемы  
Выход – статистики  $\mu, \sigma$  для каждой фонемы
  1. Duration predictor как в FastSpeech2 (свертки+линейный слой)  
Для gt значений не используется внешняя модель  
Используется алгоритм динамического программирования  
**Monotonic alignment search** (алгоритм Витерби)

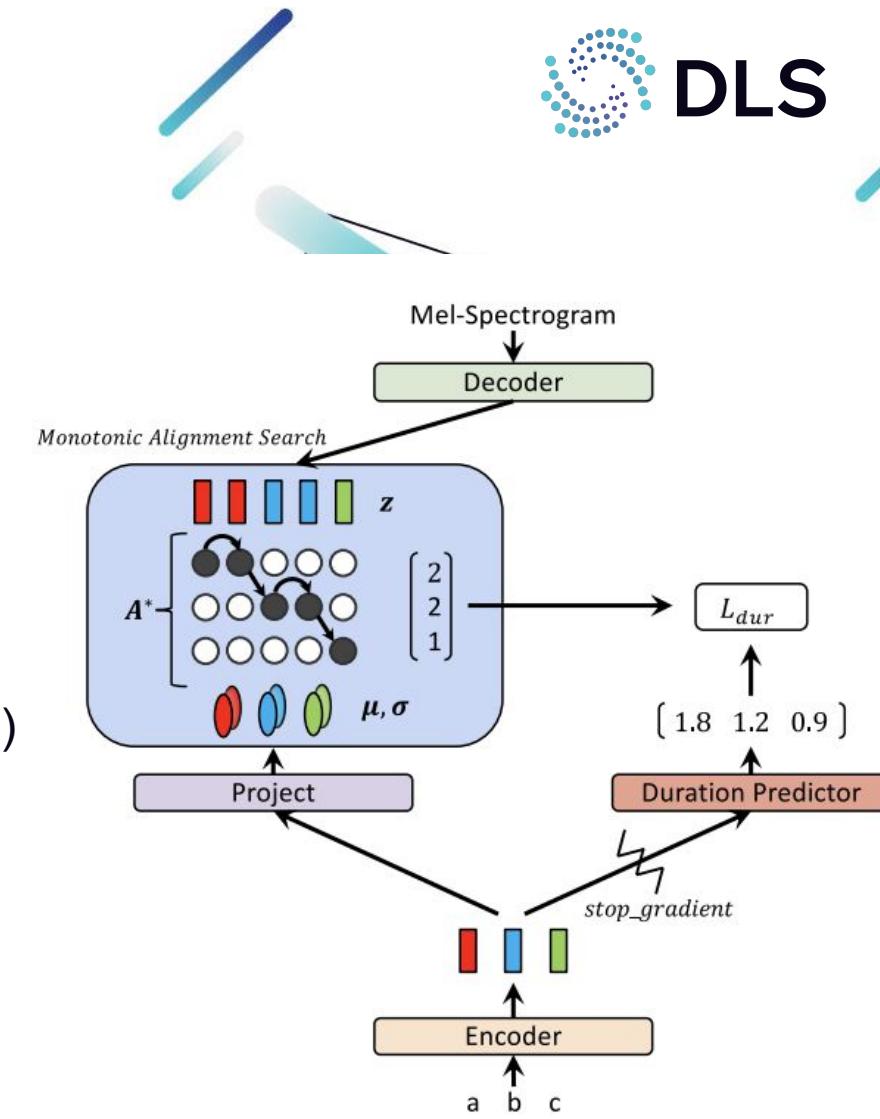


(a) An abstract diagram of the training procedure.

[Статья](#), [github](#)

# Glow-TTS (2020)

- ❑ Flow-based модель
- ❑ 3 основных блока:
  1. Encoder: Transformer, фонемы  
Выход – статистики  $\mu, \sigma (=1)$  для каждой фонемы
  1. Duration predictor как в FastSpeech2 (свертки+линейный слой)  
Для gt значений не используется внешняя модель  
Используется алгоритм динамического программирования  
Monotonic alignment search (алгоритм Витерби)
  3. Decoder: flow-based, преобразует мел-спектрограммы в латентное пространство при обучении и наоборот при инференсе



(a) An abstract diagram of the training procedure.

[Статья](#), [github](#)

# Glow-TTS (2020)

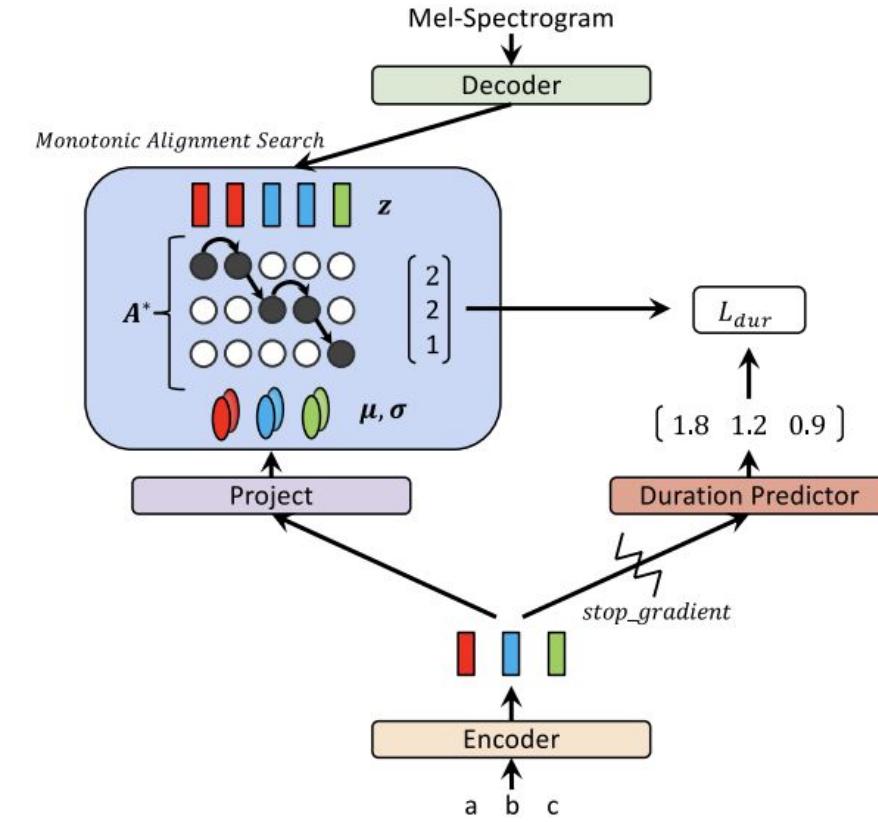
## □ Loss

- $x$  – спектрограмма,  $c$  – текст,  $z$  – латентное пространство
- log-likelihood  $\log P_X(x|c) = \log P_Z(z|c) + \log \left| \det \frac{\partial f_{dec}^{-1}(x)}{\partial x} \right|$
- априорное распределение  $P_Z$  – изотропное многомерное гауссовское распределение, параметризуется параметрами сети  $\theta$  и выравниванием  $A$

$$\log P_Z(z|c; \theta, A) = \sum_{j=1}^{T_{mel}} \log \mathcal{N}(z_j; \mu_{A(j)}, \sigma_{A(j)}).$$

$$\max_{\theta, A} L(\theta, A) = \max_{\theta, A} \log P_X(x|c; A, \theta)$$

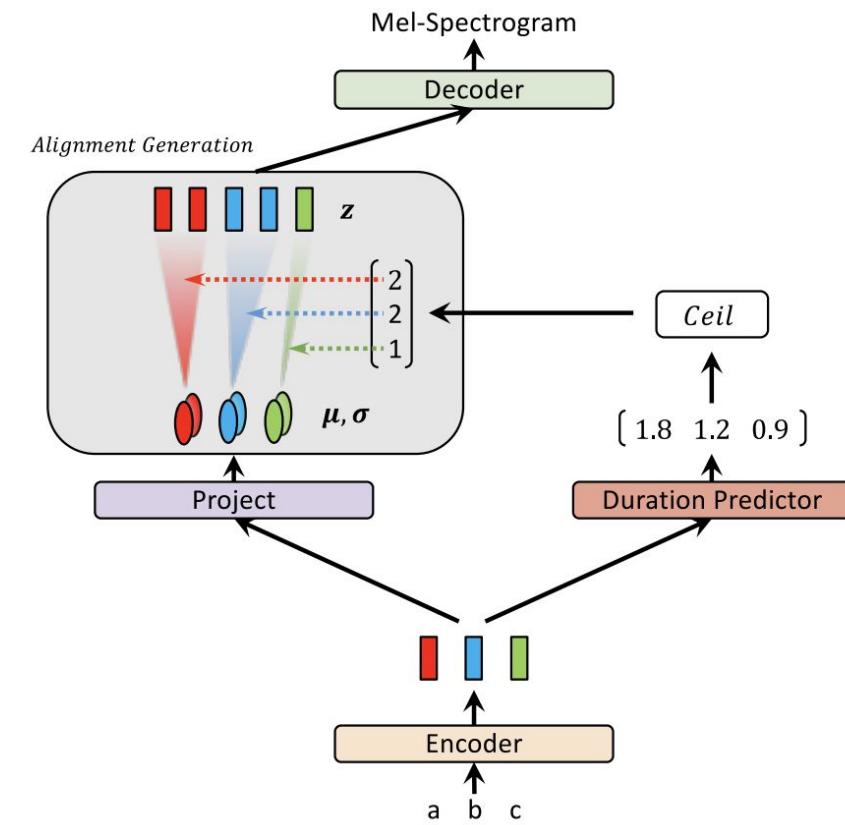
- Итеративное обновление  $\theta$  затем  $A$
- MSE loss на длительности



[Статья](#), [github](#)

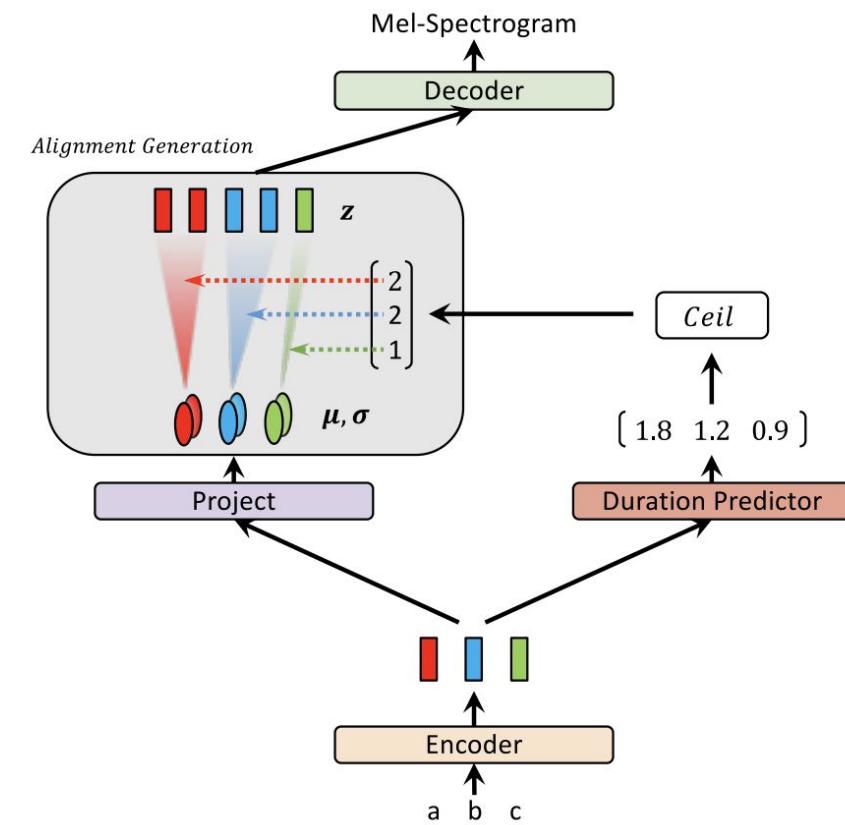
# Glow-TTS (2020)

- ❑ На инференсе используется обученный DP
- ❑ Данные:
  - 24 часа, один женский голос LJSpeech
  - 54 часа, 247 голос, LibriTTS test-clean-100
- ❑ Ресурсы:
  - single-speaker 2 GPU, 3 дня
  - multi-speaker 4GPU (speaker embedding)



# Glow-TTS (2020)

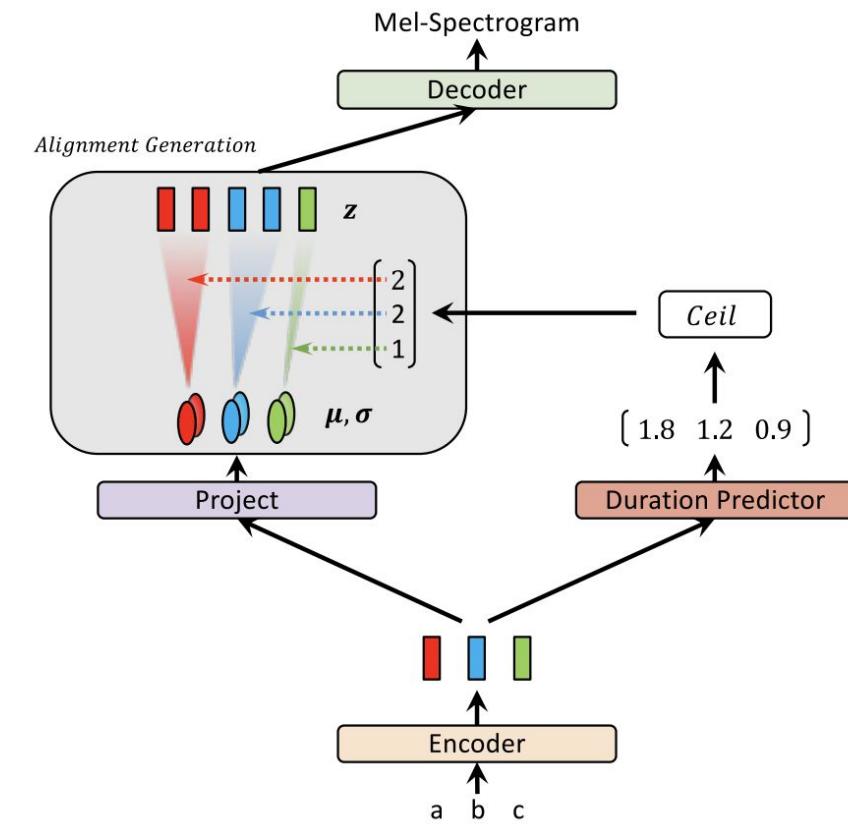
- ❑ На инференсе используется обученный DP
- ❑ Данные:
  - 24 часа, один женский голос LJSpeech
  - 54 часа, 247 голос, LibriTTS test-clean-100
- ❑ Ресурсы:
  - single-speaker 2 GPU, 3 дня
  - multi-speaker 4GPU (speaker embedding)
- ❑ Размер: 28.6M
- ❑ RTF: 0.01
- ❑ Вокодер: WaveGlow



# Glow-TTS (2020)

- ❑ На инференсе используется обученный DP
- ❑ Данные:
  - 24 часа, один женский голос LJSpeech
  - 54 часа, 247 голос, LibriTTS test-clean-100
- ❑ Ресурсы:
  - single-speaker 2 GPU, 3 дня
  - multi-speaker 4GPU (speaker embedding)
- ❑ Размер: 28.6M
- ❑ RTF: 0.01
- ❑ Вокодер: WaveGlow

single-speaker	GT		Glow-TTS	
Method				9-scale MOS
GT				$4.54 \pm 0.06$
GT (Mel + WaveGlow)				$4.19 \pm 0.07$
Tacotron2 (Mel + WaveGlow)				$3.88 \pm 0.08$
Glow-TTS ( $T = 0.333$ , Mel + WaveGlow)				$4.01 \pm 0.08$
Glow-TTS ( $T = 0.500$ , Mel + WaveGlow)				$3.96 \pm 0.08$
Glow-TTS ( $T = 0.667$ , Mel + WaveGlow)				$3.97 \pm 0.08$



multi-speaker	GT		Glow-TTS	
Method				9-scale MOS
GT				$4.54 \pm 0.07$
GT (Mel + WaveGlow)				$4.22 \pm 0.07$
Tacotron2 (Mel + WaveGlow)				$3.35 \pm 0.12$
Glow-TTS ( $T = 0.333$ , Mel + WaveGlow)				$3.20 \pm 0.12$
Glow-TTS ( $T = 0.500$ , Mel + WaveGlow)				$3.31 \pm 0.12$
Glow-TTS ( $T = 0.667$ , Mel + WaveGlow)				$3.45 \pm 0.11$

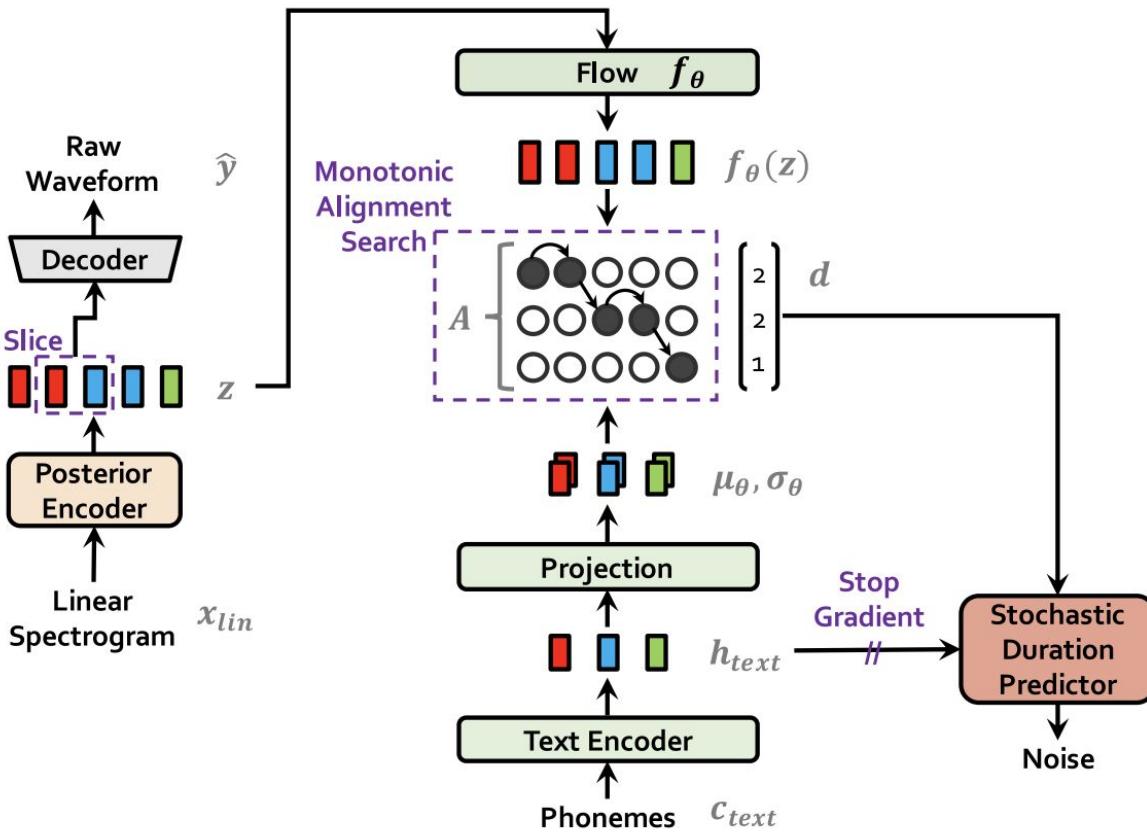
[Demo](#)

# Glow-TTS (2020)



- Быстрая (неавторегрессионная)
- Не требует внешней модели для извлечения длительностей
- Разнообразная речь, можно менять температуру
- Контроль темпом речи

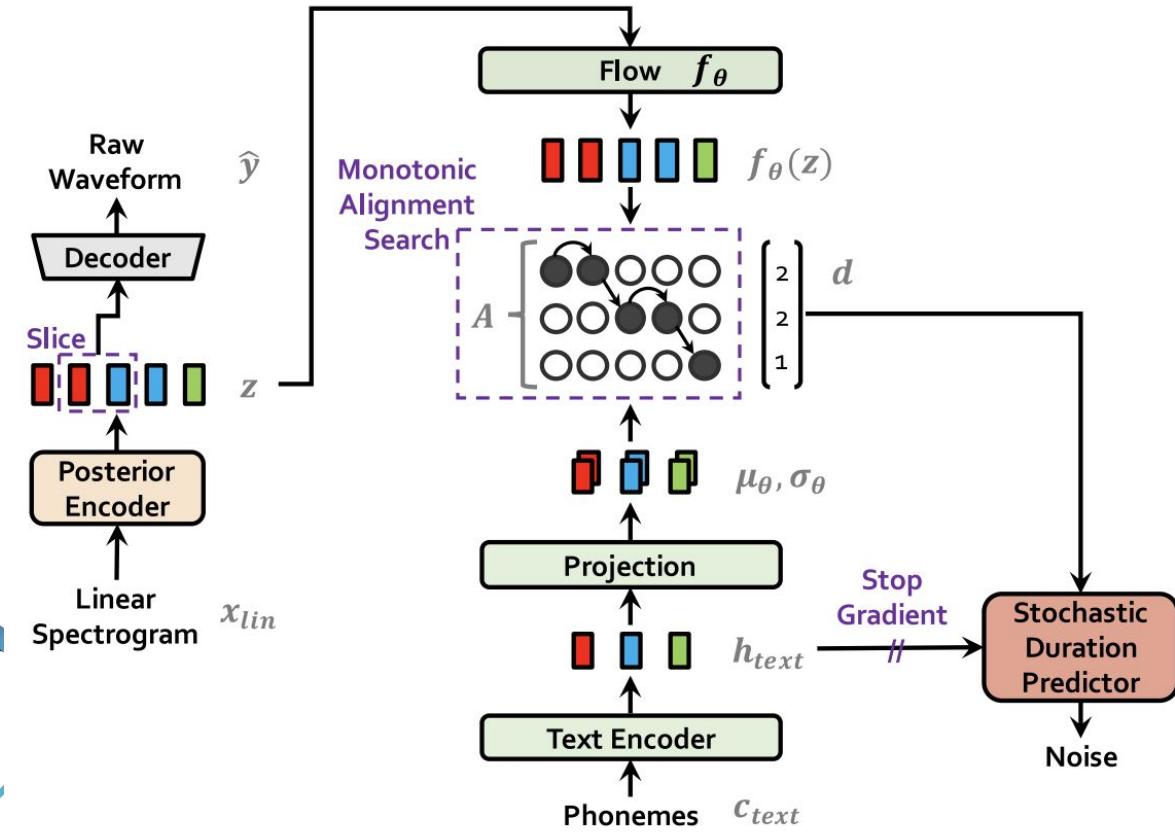
# VITS (2021)



- End-to-end model
- Объединение моделей Glow-TTS и HiFi-GAN
- Дополнительные adversarial лоссы
- Модификация Duration Predictor
- Данные:
  - 24 часа, один женский голос LJSpeech
  - 44 часа, 109 голосов, VCTK
- Ресурсы: 4 GPU

[Статья](#), [github](#)

# VITS (2021)



[Статья](#), [github](#)

single-speaker

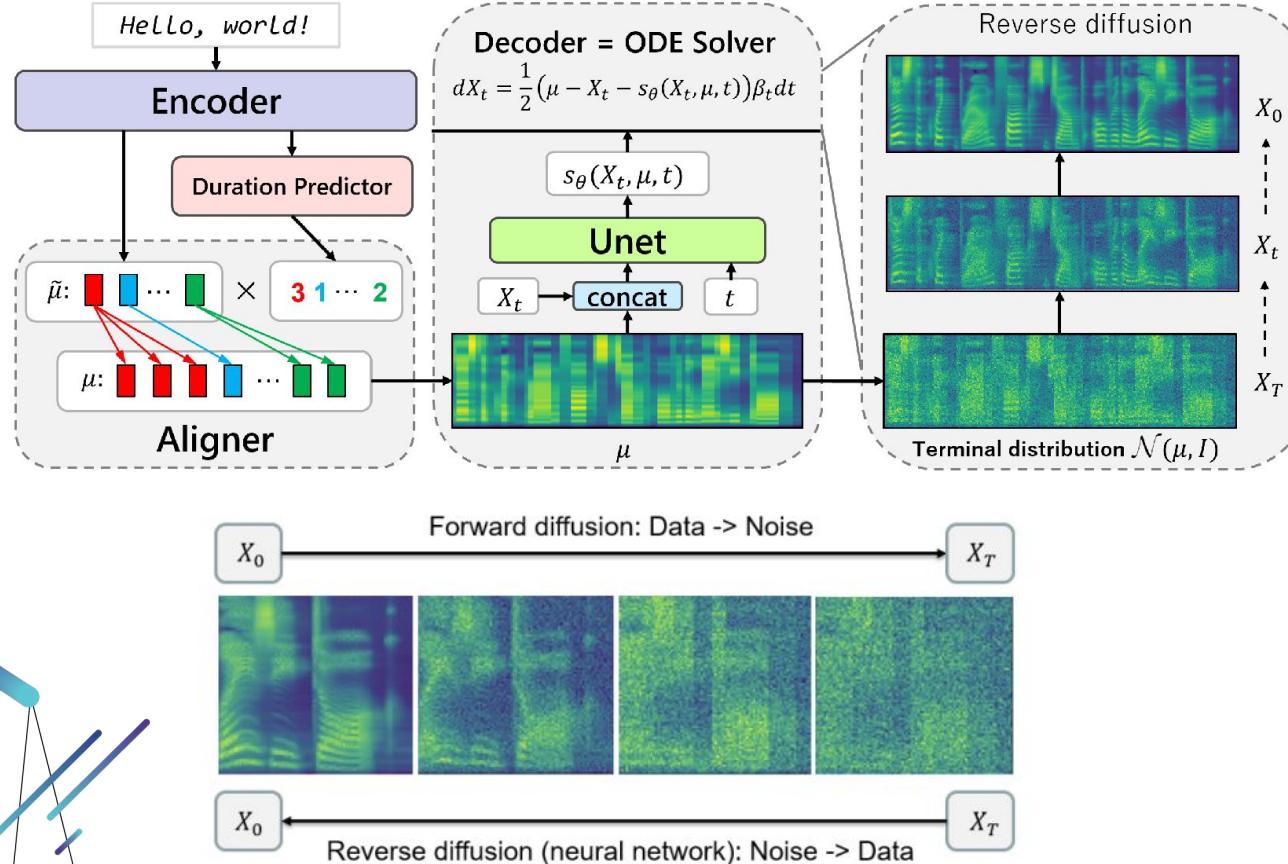
Model	MOS (CI)
Ground Truth	4.46 ( $\pm 0.06$ )
Tacotron 2 + HiFi-GAN	3.77 ( $\pm 0.08$ )
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 ( $\pm 0.07$ )
Glow-TTS + HiFi-GAN	4.14 ( $\pm 0.07$ )
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 ( $\pm 0.07$ )
VITS (DDP)	4.39 ( $\pm 0.06$ )
<b>VITS</b>	<b>4.43 (<math>\pm 0.06</math>)</b>

multi-speaker

Model	MOS (CI)
Ground Truth	4.38 ( $\pm 0.07$ )
Tacotron 2 + HiFi-GAN	3.14 ( $\pm 0.09$ )
Tacotron 2 + HiFi-GAN (Fine-tuned)	3.19 ( $\pm 0.09$ )
Glow-TTS + HiFi-GAN	3.76 ( $\pm 0.07$ )
Glow-TTS + HiFi-GAN (Fine-tuned)	3.82 ( $\pm 0.07$ )
<b>VITS</b>	<b>4.38 (<math>\pm 0.06</math>)</b>

[Demo](#)

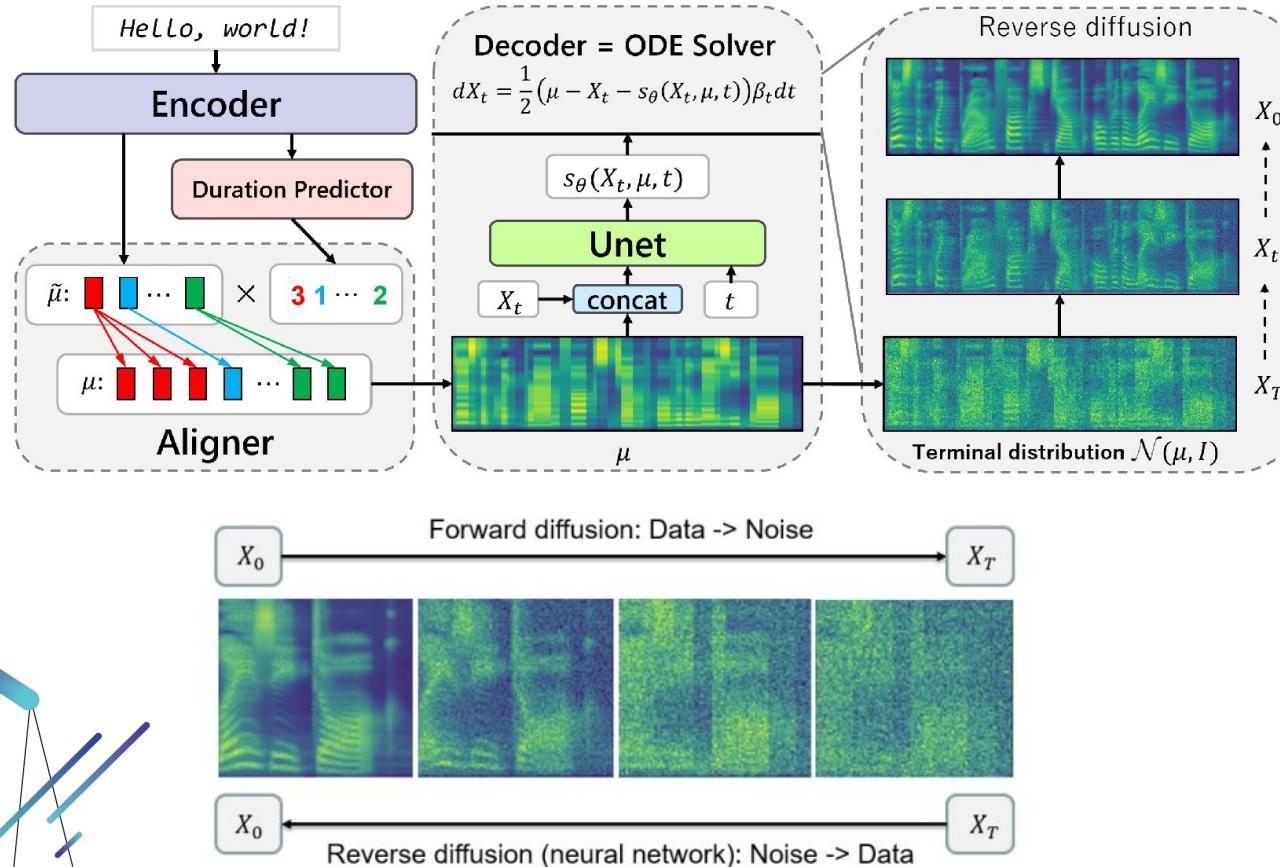
# Grad-TTS (2021)



- Encoder и duration predictor и алгоритм выравнивания MAS из Glow-TTS
- Основное нововведение – диффузионный декодер
- Параметризуется в непрерывном виде с помощью ОДУ
- Декодер обуславливается на выходы энкодера

[Статья](#), [github](#)

# Grad-TTS (2021)



[Статья](#), [github](#)

- Ресурсы: 1 GPU
- Данные: 24 часа, один женский голос LJSpeech
- Размер: 14.8M
- RTF: ~0.03
- Вокодер: HiFi-GAN
- 10 итераций на инференсе

Table 1. Model comparison.

Model	Enc params <sup>†</sup>	Dec params	RTF	Log-likelihood	MOS
Grad-TTS-1000	7.2m	7.6m	3.663	<b>0.174 ± 0.001</b>	<b>4.44 ± 0.05</b>
Grad-TTS-100			0.363		4.38 ± 0.06
Grad-TTS-10			0.033		4.38 ± 0.06
Grad-TTS-4			0.012		3.96 ± 0.07
Glow-TTS	7.2m	21.4m	0.008	0.082	4.11 ± 0.07
FastSpeech	24.5m		<b>0.004</b>	—	3.68 ± 0.09
Tacotron2	28.2m		0.075	—	4.32 ± 0.07
Ground Truth	—	—	—	—	4.53 ± 0.06

[Demo](#)

# Вокодеры

# WaveNet (2016)

- ❑ end-to-end генерация/замена вокодера в параметрической системе
- ❑ Авторегрессионная модель
- ❑ CNN (1D):
  - causal convolution
  - dilated convolution
- ❑ Стэк таких блоков с residual connection
- ❑ Выход – softmax, распределение на 256 квантизованных значения
- ❑ Частота семплирования 16kHz

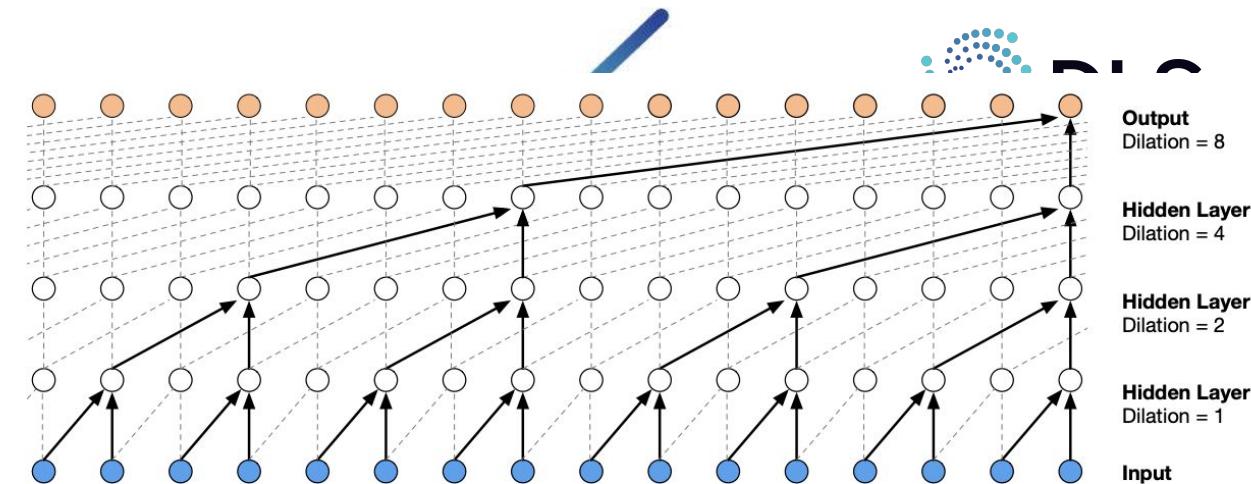
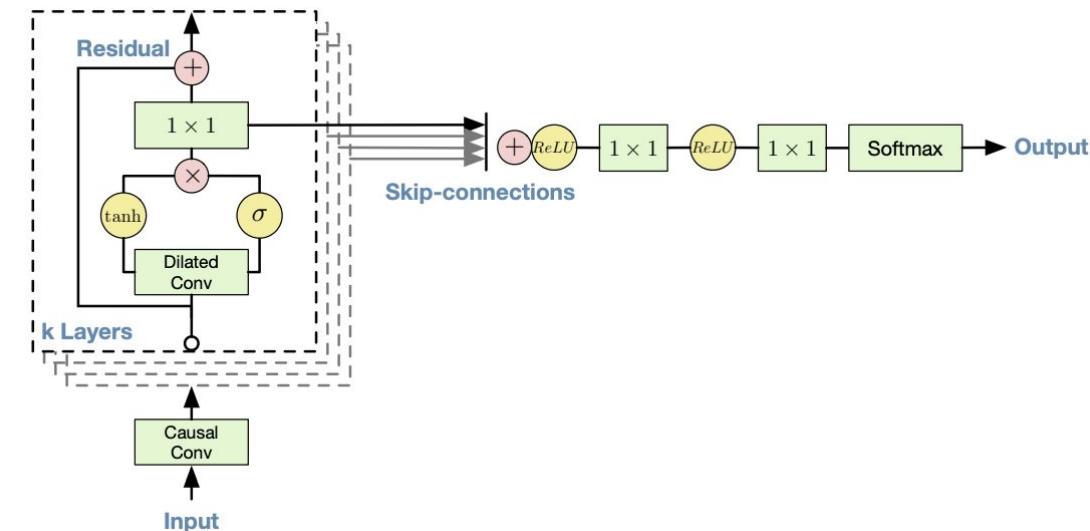


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.



[Paper](#)

# WaveNet (2016)

Варианты тестирования:

1. Unconditional generations (one-hot speaker vector)

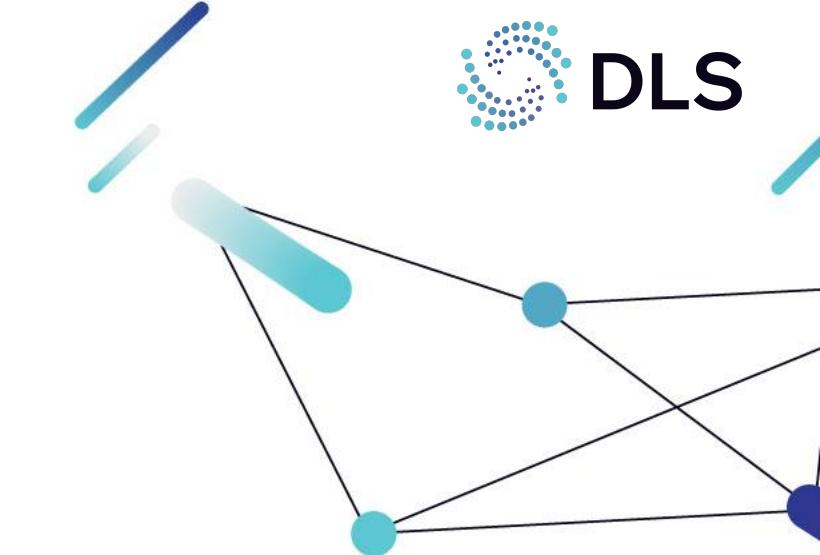
- обучение на датасете VCTK (44 часа, 109 спикеров)
- проверяется общая интонация и возможность контроля голосом

1. TTS

- 24.6 часов английской речи, 1 женский голос
- 34.8 часов китайской речи, 1 женский голос
- обуславливаются на лингвистические признаки, log F0, durations

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	$3.67 \pm 0.098$	$3.79 \pm 0.084$
HMM-driven concatenative	$3.86 \pm 0.137$	$3.47 \pm 0.108$
<b>WaveNet (L+F)</b>	<b><math>4.21 \pm 0.081</math></b>	<b><math>4.08 \pm 0.085</math></b>
Natural (8-bit $\mu$ -law)	$4.46 \pm 0.067$	$4.25 \pm 0.082$
Natural (16-bit linear PCM)	$4.55 \pm 0.075$	$4.21 \pm 0.071$

3. Music generation



Parametric

Concatenative

WaveNet

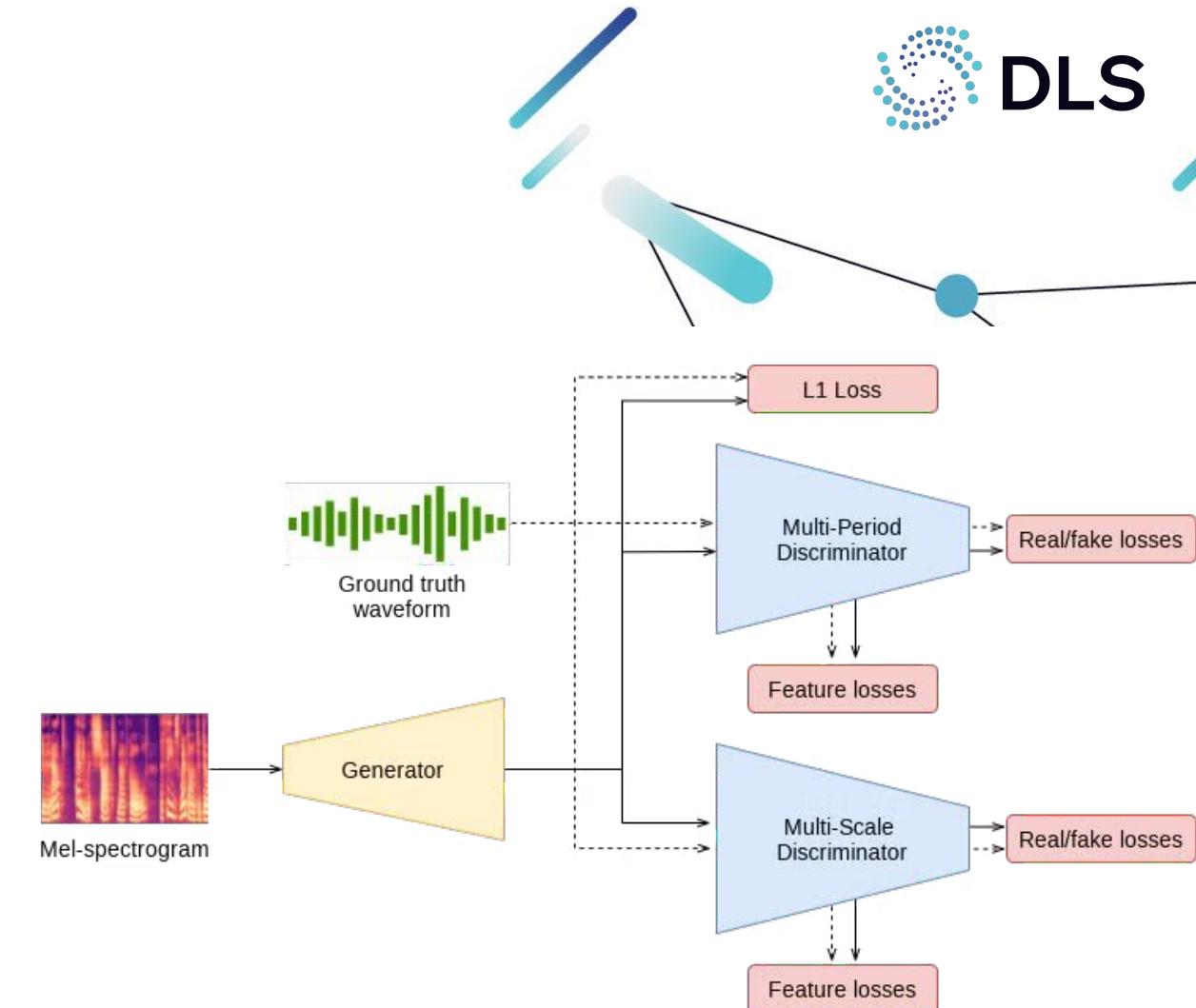
WaveNet music

[Demo](#)

# HiFi-GAN (2020)

Вокодер: мел-спектrogramma → аудио

GAN: генератор + 2 дискриминатора



картишка позаимствована [отсюда](#)

[Paper](#), [github](#)

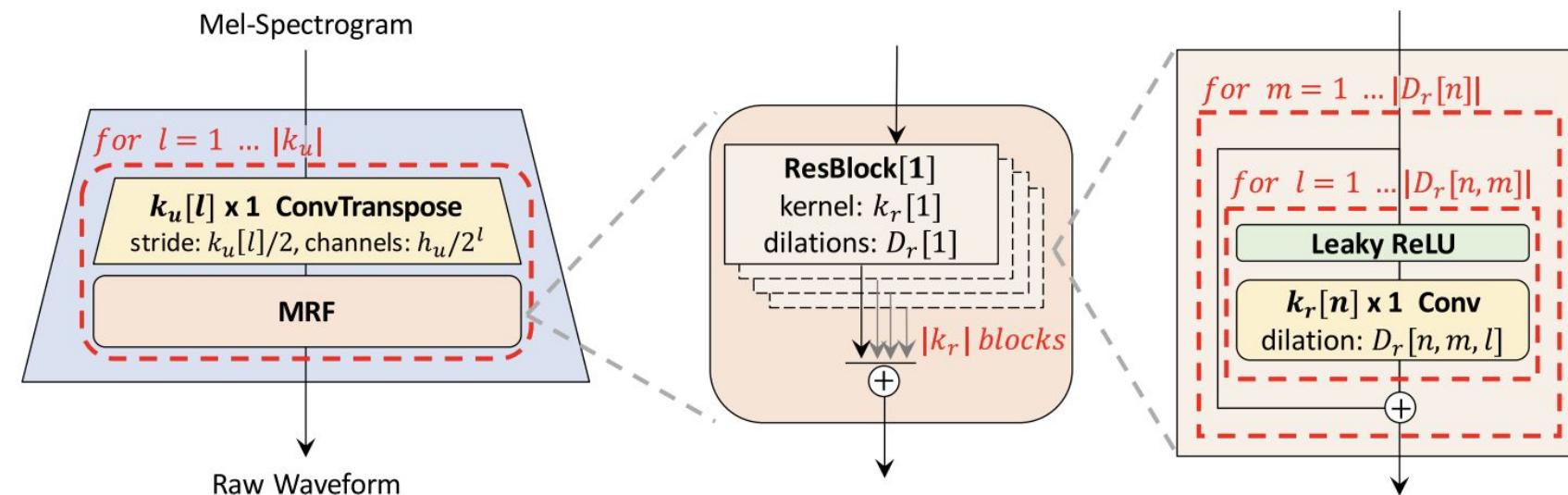
# HiFi-GAN (2020)

Вокодер: мел-спектrogramma → аудио

GAN: генератор + 2 дискриминатора

**Генератор:** upsampling из количества фреймов в количество сэмплов

- CNN, состоит из:
  - transposed convolution (upsampling)
  - MRF (multi-receptive field fusion): residual blocks со свертками с разной длиной ядер и dilation
- не принимает на вход шум



# HiFi-GAN (2020)

Вокодер: мел-спектrogramma → аудио

GAN: генератор + 2 дискриминатора

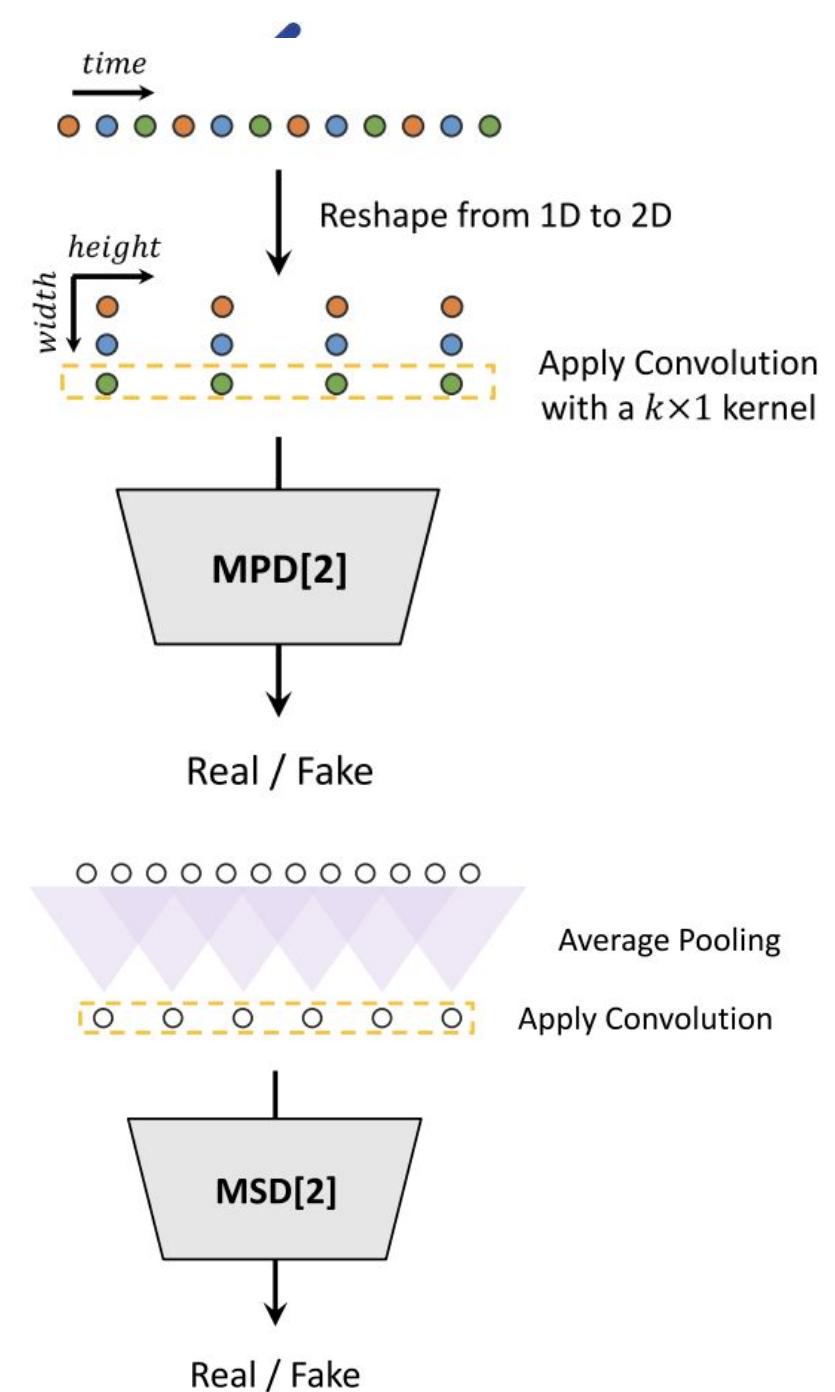
**Дискриминатор:** учесть периодичность аудио сигнала

## 1) Multi-period discriminator (MPD)

- обрабатывает сэмплы аудио с промежутками (например, каждый второй)
- periods = [2, 3, 5, 7, 11]
- CNN

## 2) Multi-Scale discriminator (MSD)

- работает с последовательностью агрегированной x2 average pooling, x4 avg pooling
- CNN



# HiFi-GAN (2020)

Вокодер: мел-спектrogramma → аудио

GAN: генератор + 2 дискриминатора

**Loss:**

- adversarial

$$\mathcal{L}_{Adv}(D; G) = \mathbb{E}_{(x,s)} \left[ (D(x) - 1)^2 + (D(G(s)))^2 \right]$$

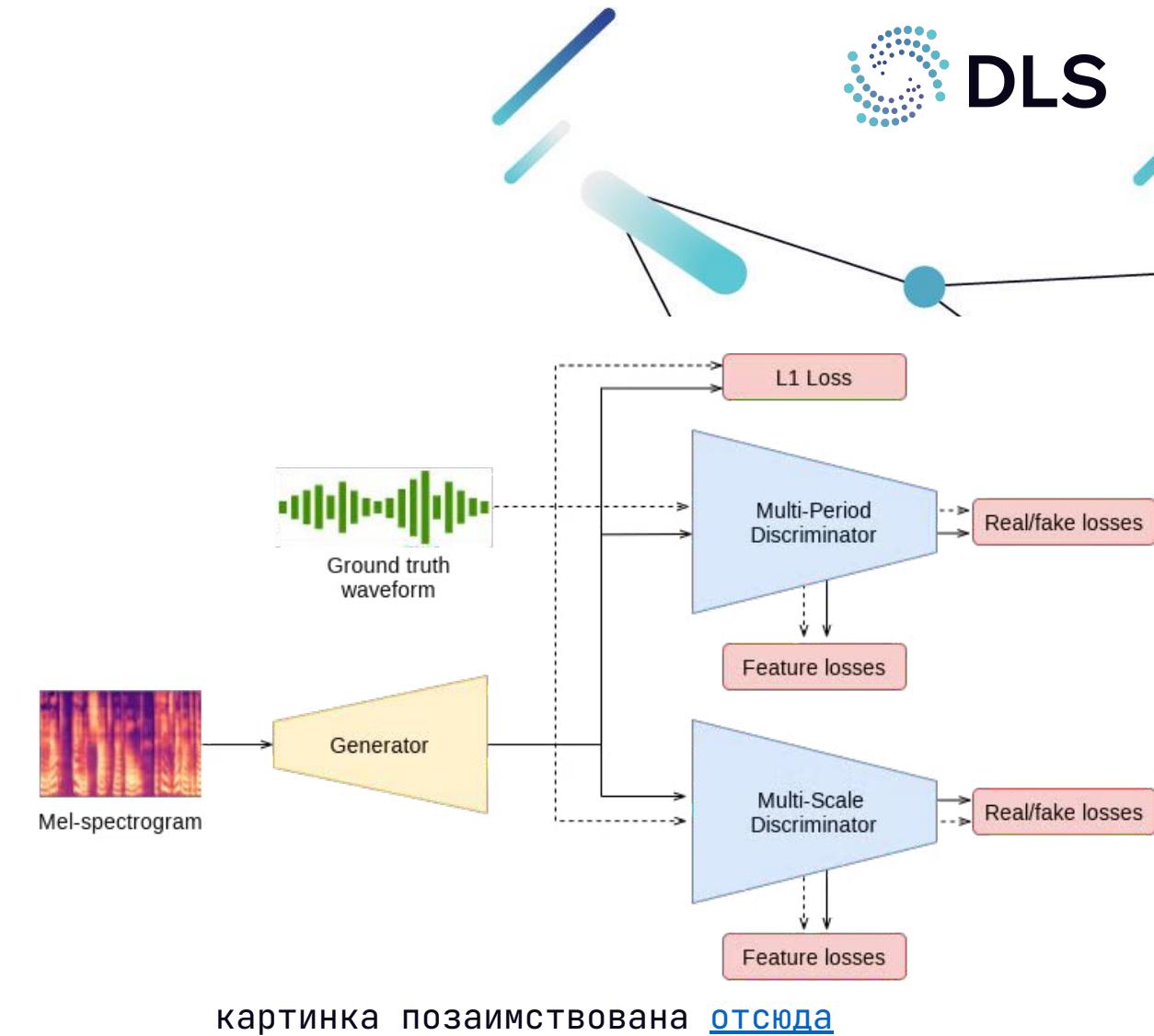
$$\mathcal{L}_{Adv}(G; D) = \mathbb{E}_s \left[ (D(G(s)) - 1)^2 \right]$$

- mel-spectrogram loss (для генератора)

$$\mathcal{L}_{Mel}(G) = \mathbb{E}_{(x,s)} \left[ \|\phi(x) - \phi(G(s))\|_1 \right]$$

- feature matching loss (для генератора)

$$\mathcal{L}_{FM}(G; D) = \mathbb{E}_{(x,s)} \left[ \sum_{i=1}^T \frac{1}{N_i} \|D^i(x) - D^i(G(s))\|_1 \right]$$



картина позаимствована [отсюда](#)

[Paper](#), [github](#)

# HiFi-GAN (2020)

- ❑ Ресурсы: 1 GPU
- ❑ Данные:
  - 24 часа, один женский голос LJSpeech
  - 44 часа, 109 голосов, VCTK
- ❑ Частота семплирования 22kHz

Model	MOS (CI)	Speed on CPU (kHz)	Speed on GPU (kHz)	# Param (M)
Ground Truth	4.45 ( $\pm 0.06$ )	—	—	—
WaveNet (MoL)	4.02 ( $\pm 0.08$ )	—	0.07 ( $\times 0.003$ )	24.73
WaveGlow	3.81 ( $\pm 0.08$ )	4.72 ( $\times 0.21$ )	501 ( $\times 22.75$ )	87.73
MelGAN	3.79 ( $\pm 0.09$ )	145.52 ( $\times 6.59$ )	14,238 ( $\times 645.73$ )	4.26
HiFi-GAN V1	<b>4.36</b> ( $\pm 0.07$ )	31.74 ( $\times 1.43$ )	3,701 ( $\times 167.86$ )	13.92
HiFi-GAN V2	4.23 ( $\pm 0.07$ )	214.97 ( $\times 9.74$ )	16,863 ( $\times 764.80$ )	<b>0.92</b>
HiFi-GAN V3	4.05 ( $\pm 0.08$ )	<b>296.38</b> ( $\times 13.44$ )	<b>26,169</b> ( $\times 1,186.80$ )	1.46

# HiFi-GAN (2020)

Unseen speakers

Table 3: Quality comparison of synthesized utterances for unseen speakers.

Model	MOS (CI)
Ground Truth	3.79 ( $\pm 0.07$ )
WaveNet (MoL)	3.52 ( $\pm 0.08$ )
WaveGlow	3.52 ( $\pm 0.08$ )
MelGAN	3.50 ( $\pm 0.08$ )
HiFi-GAN V1	<b>3.77</b> ( $\pm 0.07$ )
HiFi-GAN V2	3.69 ( $\pm 0.07$ )
HiFi-GAN V3	3.61 ( $\pm 0.07$ )

Tacotron2 + HiFi-GAN

Table 4: Quality comparison for end-to-end speech synthesis.

Model	MOS (CI)
Ground Truth	4.23 ( $\pm 0.07$ )
WaveGlow (w/o fine-tuning)	3.69 ( $\pm 0.08$ )
HiFi-GAN V1 (w/o fine-tuning)	3.91 ( $\pm 0.08$ )
HiFi-GAN V2 (w/o fine-tuning)	3.88 ( $\pm 0.08$ )
HiFi-GAN V3 (w/o fine-tuning)	3.89 ( $\pm 0.08$ )
WaveGlow (find-tuned)	3.66 ( $\pm 0.08$ )
HiFi-GAN V1 (find-tuned)	<b>4.18</b> ( $\pm 0.08$ )
HiFi-GAN V2 (find-tuned)	4.12 ( $\pm 0.07$ )
HiFi-GAN V3 (find-tuned)	4.02 ( $\pm 0.08$ )



# Спасибо за внимание!

sadekova.t.r@gmail.com

Садекова Таснима

Huawei

