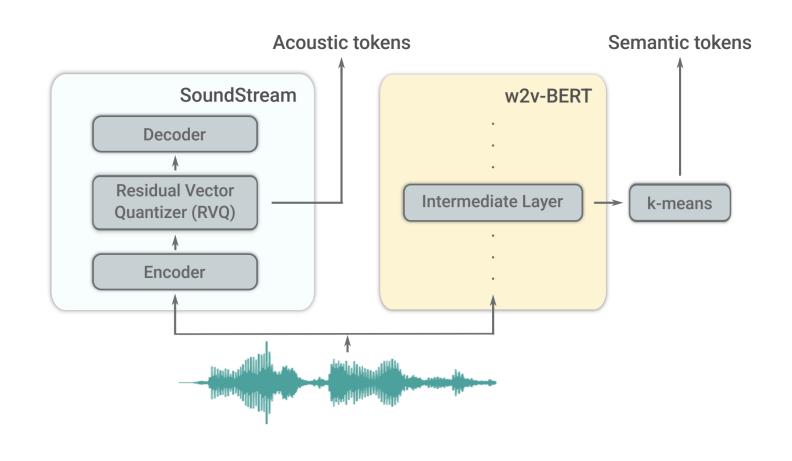
# SoundStorm: Efficient Parallel Audio Generation

Бекян Артём

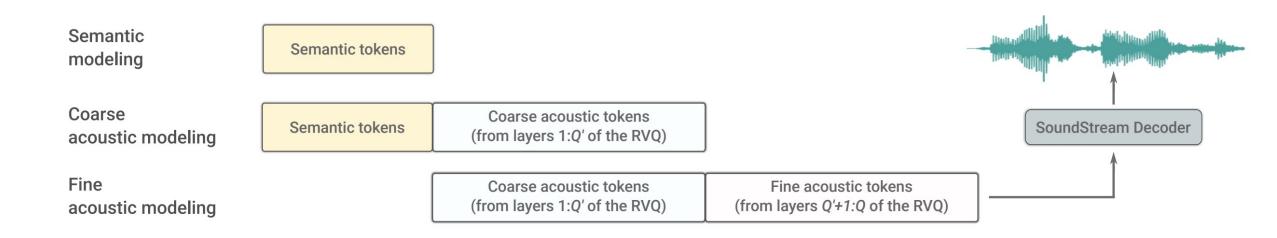
#### Предисловие. AudioLM и его токены

- Semantic tokens хорошо захватывают данные о глобальной структуре, но плохо о локальной
- Acoustic tokens захватывают акустические локальные детали аудио, чтобы генерация была более высокого качества



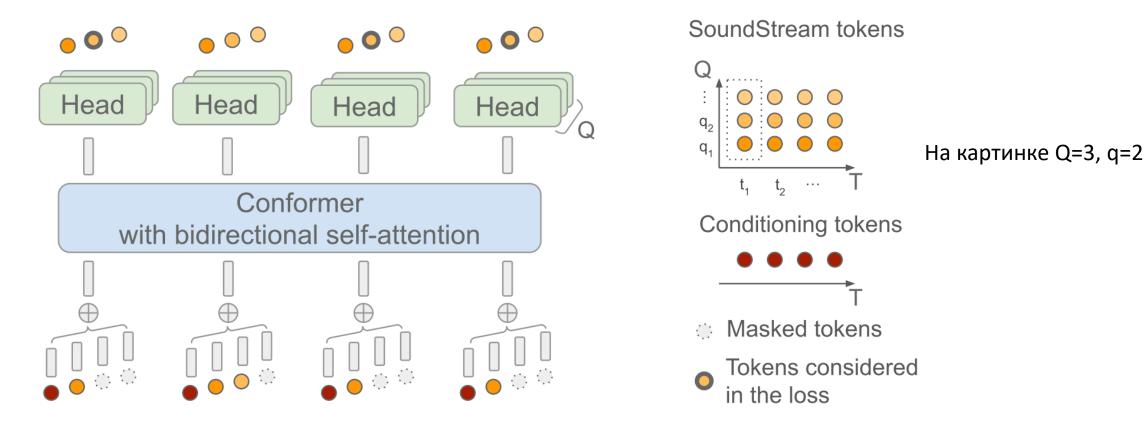
#### Предисловие. A куда SoundStorm?

SoundStorm предлагается использовать вместо *Coarse acoustic* model и Fine acoustic model



#### Архитектура

Предсказываются RVQ токены, выбирается q, все токены q+1 ... Q маскируются, некоторые токены q тоже маскируются, чтобы по ним считать лосс



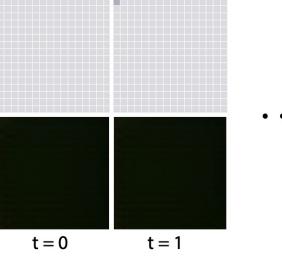
#### Архитектура. Декодинг в MaskGIT

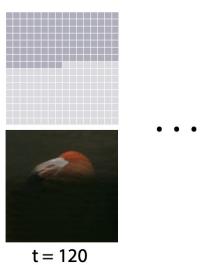
Изначально все токены маскированы. Алгоритм декодирования на итерации t:

- 1. Для каждого маскированного токена предсказываем вероятность  $p^{(t)} \in \mathbb{R}^{N imes K}$  параллельно.
- 2. Семплируем токен, основываясь на вероятностях, а величину  $p_i^{(t)}$  считаем уверенностью предсказания
- 3. В соответствии с функцией  $\gamma$  (в статье это косинус) маскируем обратно  $n = \left[\gamma\left(\frac{t}{T}\right)N\right]$ , где T общее количество итераций, при помощи сортировки уверенностей по убыванию

#### Архитектура. Декодинг в MaskGIT

Sequential Decoding with Autoregressive Transformers

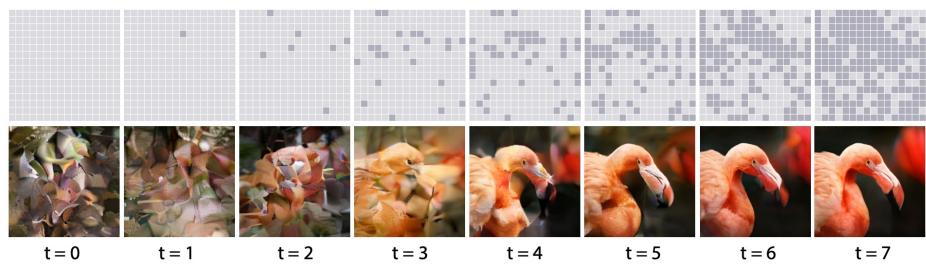








Scheduled
Parallel
Decoding
with MaskGIT



#### Архитектура. Маскирование

#### Для обучения:

- 1. (Если хотим использовать промпты) Равновероятно семплируем таймстеп для промпта от 0 до Т-1
- 2. Семплируем q (уровень RVQ) равновероятно от 1 до Q
- 3. Семплируем маску по косинусному расписанию, то есть  $u \sim \mathcal{U}[0,\pi/2], \ \ p = \cos(u), \ \ M_i = Bernoulli(p)$
- 4. Маскируем не промптовые токены, которые выбрали или которые на >q уровнях RVQ

### Архитектура. Параллельный декодинг

По сути такой же, как в MaskGIT, только использующийся для каждого уровня RVQ с 1 до Q (coarse-to-fine)

**Отличие:** при последнем проходе на каждом уровне RVQ используется жадное декодирование вместо confidence-based семплинга

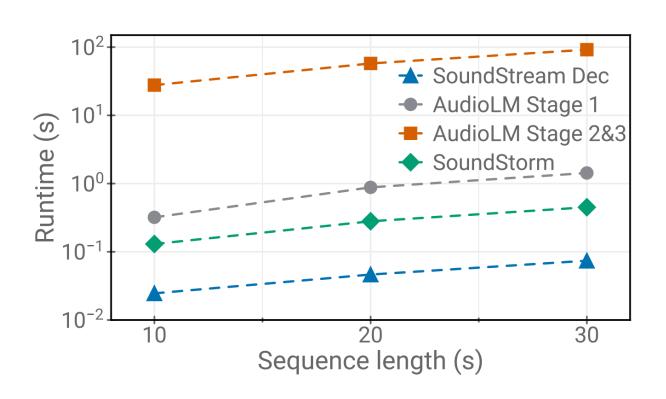
**Интересный момент:** поскольку чем дальше в RVQ, тем локальнее становятся токены, то со временем Т уменьшается

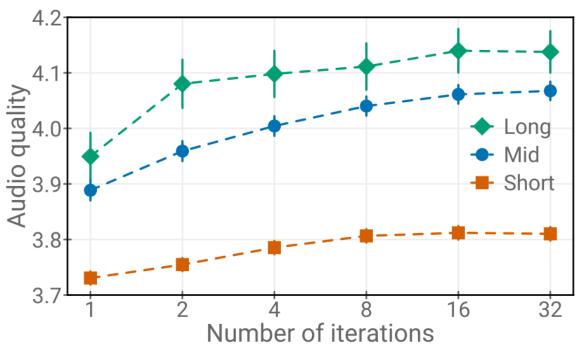
## Результаты

Рассчитаны с помощью ASR модели						MOS			Одинаковость акустически свойств						
WER↓				CER↓			Audio quality <sup>↑</sup>			Voice preservation			Acoustic consistency		
short	mid	long	short	mid	long	short	mid	long	short	mid	long	short	mid	long	
2.62	1.95	2.20	0.89	0.55	0.69	3.72	3.91	3.99	0.63	0.65	0.66	0.97	0.95	0.93	
					W	ithout a	speake	er prom	pt						
4.65	3.59	4.79	2.15	1.57	2.30	3.93	4.04	4.08	_	_	_	_	_	_	
3.48	2.55	3.33	1.39	0.89	1.29	4.01	4.16	4.20	_	_	_	_	_	_	
						With a s	speaker	prompt							
3.77 <b>2.99</b>	3.40 <b>2.43</b>	3.75 <b>3.36</b>	1.50 <b>1.10</b>	1.47 <b>0.81</b>	1.54 <b>1.24</b>	<b>3.91</b> 3.81	4.06 4.05	4.10 4.15	0.46 0.57	0.48	0.48	0.96 0.96	0.91 <b>0.94</b>	0.86 <b>0.91</b>	
	short 2.62 4.65 3.48	short mid  2.62 1.95  4.65 3.59  3.48 2.55  3.77 3.40	WER↓         short       mid       long         2.62       1.95       2.20         4.65       3.59       4.79         3.48       2.55       3.33         3.77       3.40       3.75	WER↓         short       mid       long       short         2.62       1.95       2.20       0.89         4.65       3.59       4.79       2.15         3.48       2.55       3.33       1.39         3.77       3.40       3.75       1.50	WER↓       CER↓         short       mid       long       short       mid         2.62       1.95       2.20       0.89       0.55         4.65       3.59       4.79       2.15       1.57         3.48       2.55       3.33       1.39       0.89         3.77       3.40       3.75       1.50       1.47	WER↓       CER↓         short       mid       long       short       mid       long         2.62       1.95       2.20       0.89       0.55       0.69         4.65       3.59       4.79       2.15       1.57       2.30         3.48       2.55       3.33       1.39       0.89       1.29         3.77       3.40       3.75       1.50       1.47       1.54	WER↓       CER↓         short       mid       long       short       mid       long       short         2.62       1.95       2.20       0.89       0.55       0.69       3.72         Without of the color of the	WER↓         CER↓         Audio quality           short         mid         long         short         mid         long         short         mid           2.62         1.95         2.20         0.89         0.55         0.69         3.72         3.91           4.65         3.59         4.79         2.15         1.57         2.30         3.93         4.04           3.48         2.55         3.33         1.39         0.89         1.29         4.01         4.16           With a speaker         3.77         3.40         3.75         1.50         1.47         1.54         3.91         4.06	WER↓         CER↓         Audio quality           short         mid         long         short         mid         long         short         mid         long           2.62         1.95         2.20         0.89         0.55         0.69         3.72         3.91         3.99           4.65         3.59         4.79         2.15         1.57         2.30         3.93         4.04         4.08           3.48         2.55         3.33         1.39         0.89         1.29         4.01         4.16         4.20           With a speaker prompt           3.77         3.40         3.75         1.50         1.47         1.54         3.91         4.06         4.10	WER   CER   CER	WER↓         CER↓         Audio quality         Woice preservation           short         mid         long         short         long         short         long         short         long         short         long         short         long         short         long	Рассчитаны с помощью ASR модели         MOS           WER↓         CER↓         Audio quality preservation           short         mid         long         nid         long           2.62         1.95         2.20         0.89         0.69         3.72         3.91         3.99         0.63         0.65         0.66           4.65         3.59         4.79         2.15         1.57         2.30         3.93         4.04         4.08         -         -         -         -           3.48         2.55         3.33         1.39         0.89<	Рассчитаны с помощью ASR модели         MOS           WER↓         CER↓         Audio quality ↑ preservation ↑         Voice preservation ↑         A colspan="8">A colspan="8">Audio quality ↑ preservation ↑         A colspan="8">Audio quality ↑ preservation ↑         A colspan="8">Audio quality ↑ preservation ↑         Audio quality ↑ preservation ↑ preservation ↑ preservation ↑         Audio quality ↑ preservation ↑ preservatio	Рассчитаны с помощью ASR модели  WER↓	

Сохранение личности спикера Рассчитано с помощью ASV модели

### Парочка аблейшнов





Только для 1 уровня RVQ, остальные декодируются жадным алгоритмом, так лучше выходит