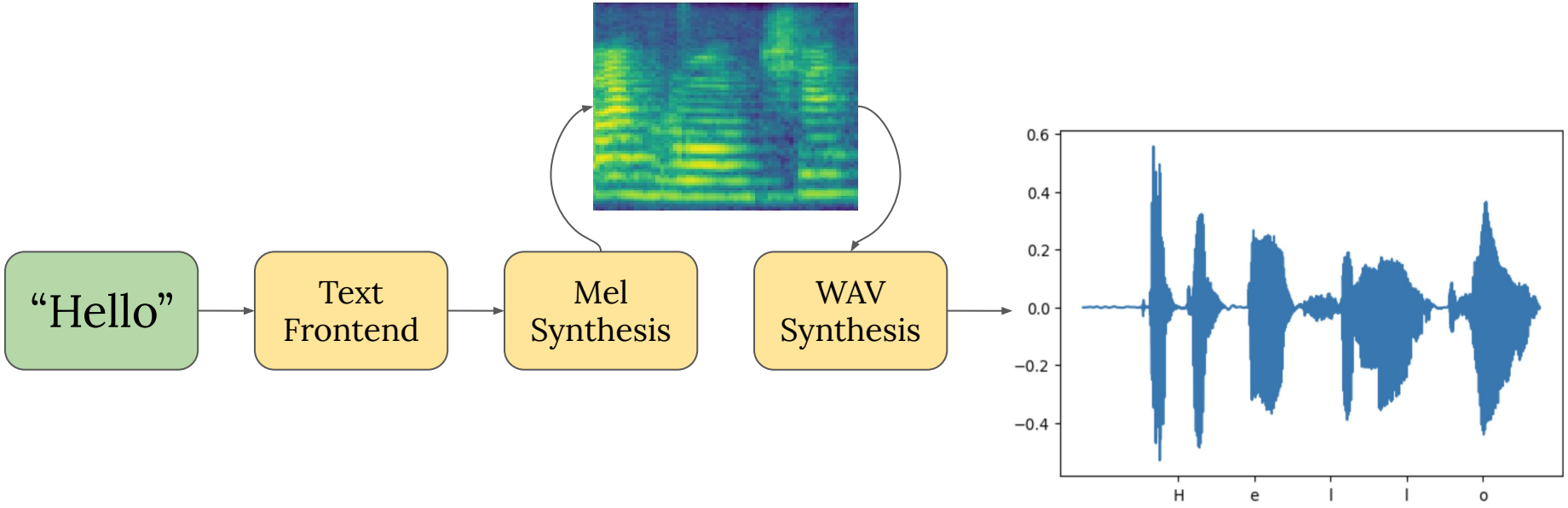
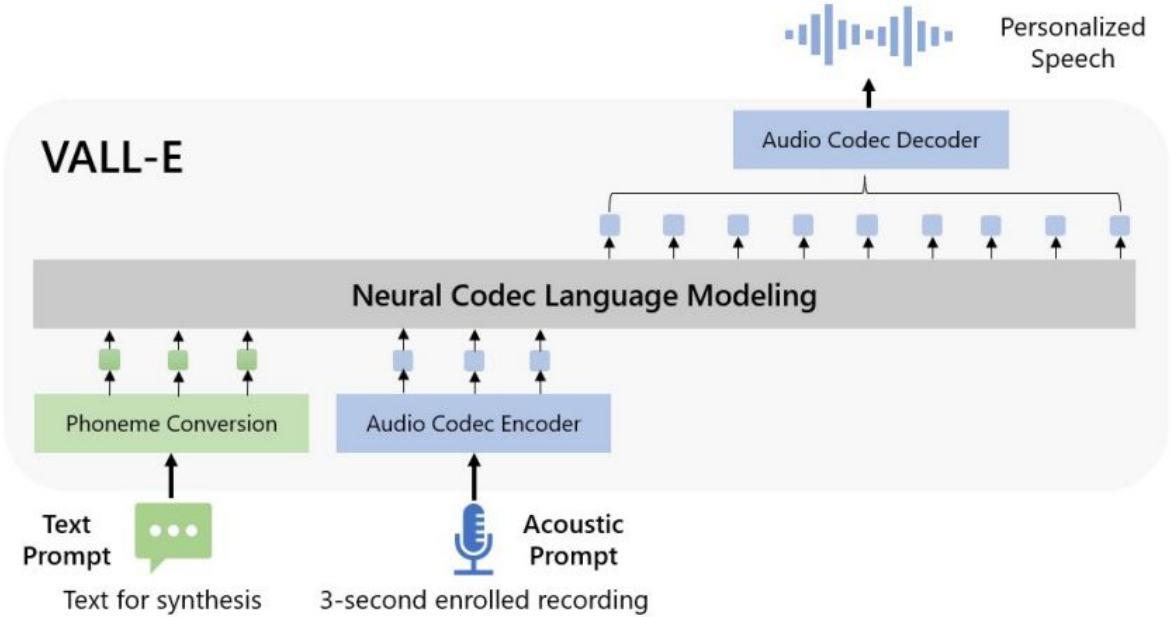
Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

Докладчик: Кокорина Юлия
Рецензент-исследователь: Василевская Юлия
Хакер: Тимонина Мария

Что было раньше



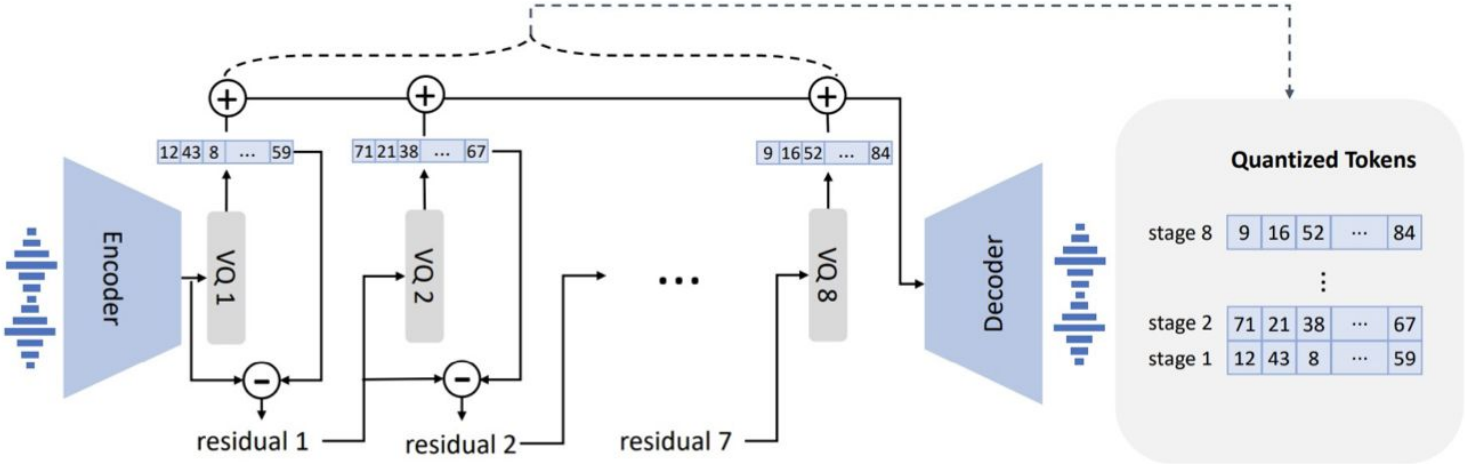
Верхнеуровневая схема



Почему нужна квантизация?

- Звук хранится как последовательность int16
- Хотим использовать генеративную модель => предсказывать вероятности
- 2^16 = 65’536, слишком много
- Будем использовать neural audio codec model

Neural audio codec model



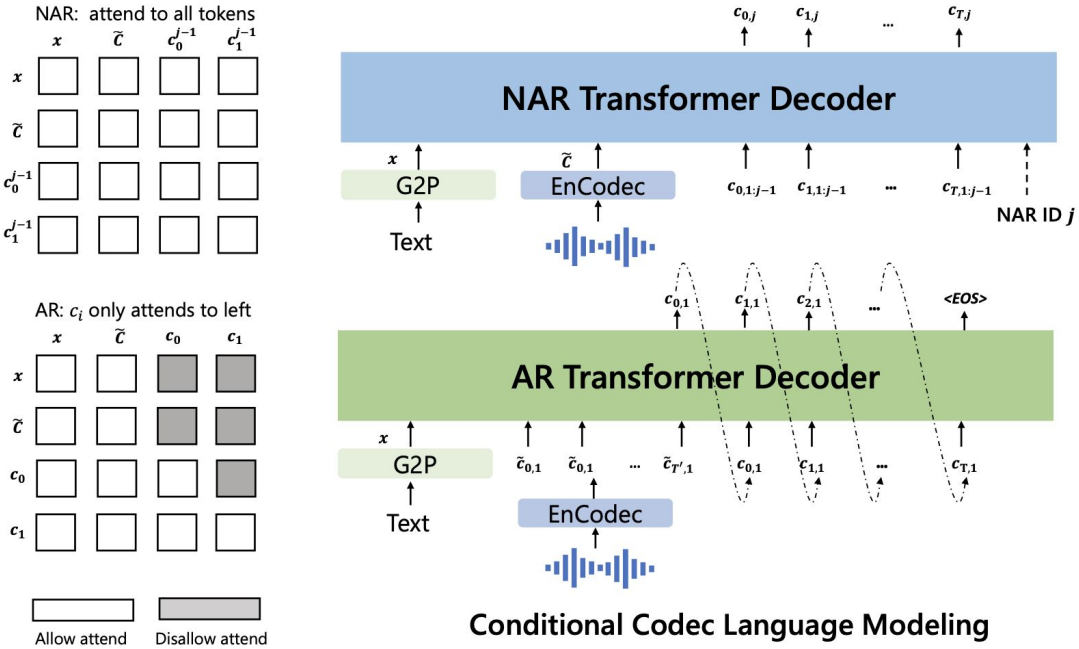
Формальная формулировка для обучения

Вход: y_i - аудио, x_i - то, что спикер говорит на этом аудио.

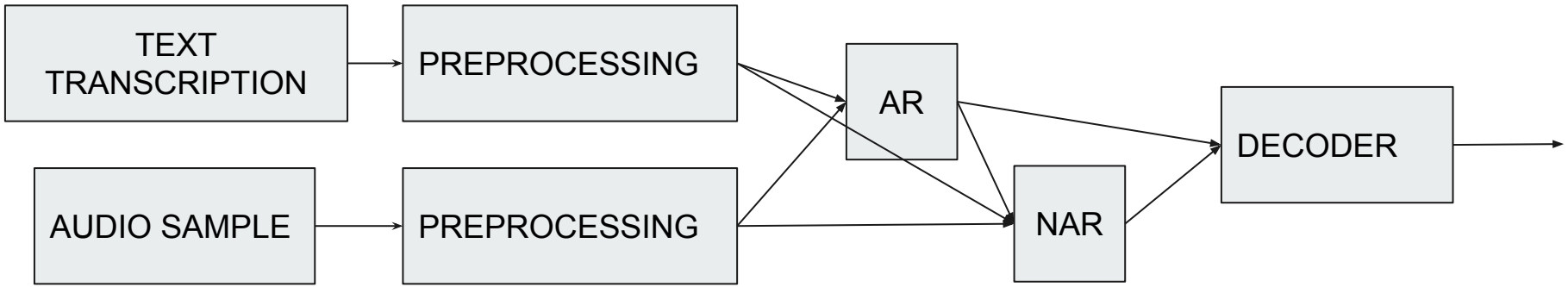
Выход: по x_i и части аудио хотим предсказать другую часть. Хотим предсказывать не звуковую волну, а акустическую матрицу

$$p(\mathbf{C}|\mathbf{x}, \tilde{\mathbf{C}}; \theta) = p(\mathbf{c}_{:,1}|\tilde{\mathbf{C}}_{:,1}, \mathbf{X}; \theta_{AR}) \prod_{j=2}^{8} p(\mathbf{c}_{:,\mathbf{j}}|\mathbf{c}_{:,<j}, \mathbf{x}, \tilde{\mathbf{C}}; \theta_{NAR})$$

2 части: AR и NAR



Inference



AR: sampling-based decoding

NAR: greedy decoding

Inference: 2 сеттинга

- VALL-E: стандартный
- VALL-E-continual: фраза, которую надо произнести - продолжение фразы из сэмпла

Experimental setup

1. Учат ASR на LibriSpeech для того, чтобы разметить LibriLight
2. Используют предобученный EnCodec
3. У AR и NAR одинаковая архитектура
4. Модель: 16 attention layers, 12 heads, embedding dim = 1024, feed-forward dim = 4096, dropout = 0.1
5. 6’000 acoustic tokens per GPU, 800’000 steps, AdamW with warm up, peak = 5 * 1e-4

Результаты: LibriSpeech

model	WER	SPK
GroundTruth	2.2	0.754
Speech-to-Speech Systems		
GSLM	12.4	0.126
AudioLM*	6.0	-
TTS Systems		
YourTTS	7.7	0.337
VALL-E	5.9	**0.580**
VALL-E-continual	**3.8**	0.508

	SMOS	CMOS (v.s. VALL-E)
YourTTS	$3.45_{\pm 0.09}$	-0.12
VALL-E	$4.38_{\pm 0.10}$	0.00
GroundTruth	$4.5_{\pm 0.10}$	+0.17

Результаты: VCTK

	3s prompt	5s prompt	10s prompt
108 full speakers			
YourTTS*	0.357	0.377	0.394
VALL-E	0.382	0.423	**0.484**
GroundTruth	0.546	0.591	0.620
11 unseen speakers			
YourTTS	0.331	0.337	0.344
VALL-E	0.389	0.380	**0.414**
GroundTruth	0.528	0.556	0.586

	SMOS	CMOS (v.s. VALL-E)
YourTTS*	$3.70_{\pm 0.09}$	-0.23
VALL-E	$3.81_{\pm 0.09}$	0.00
GroundTruth	$4.29_{\pm 0.09}$	-0.04

Результаты: несколько замечаний

- Работа модели в режиме inference не детерминирована => можем генерировать несколько вариантов
- Сохраняет звуки среды из примера спикера
- Сохраняет эмоциональную окраску речи спикера
- Иногда дублирует или не произносит некоторые слова
- Учились на датасете с аудиокнигами

Заключение

- SOtA zero-shot TTS на LibriSpeech и VCTK
- Первый раз успешно пробуют подход аналогичный LM для задач звука

Рецензия

https://arxiv.org/pdf/2301.02111.pdf

Василевская
Юля

Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

Chengyi Wang* Sanyuan Chen* Yu Wu* Ziqiang Zhang Long Zhou Shujie Liu
Zhuo Chen Yanqing Liu Huaming Wang Jinyu Li Lei He Sheng Zhao Furu Wei
Microsoft
https://github.com/microsoft/unilm

Abstract

We introduce a language modeling approach for text to speech synthesis (TTS). Specifically, we train a *neural codec language model* (called VALL-E) using discrete codes derived from an off-the-shelf neural audio codec model, and regard TTS as a conditional language modeling task rather than continuous signal regression as in previous work. During the pre-training stage, we scale up the TTS training data to 60K hours of English speech which is hundreds of times larger than existing systems. VALL-E emerges *in-context learning* capabilities and can be used to synthesize high-quality personalized speech with only a 3-second enrolled recording of an unseen speaker as an acoustic prompt. Experiment results show that VALL-E significantly outperforms the state-of-the-art zero-shot TTS system in terms of speech naturalness and speaker similarity. In addition, we find VALL-E could preserve the speaker's emotion and acoustic environment of the acoustic prompt in synthesis. See https://aka.ms/valle for demos of our work.

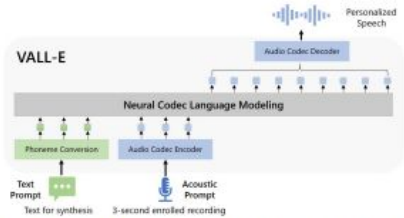
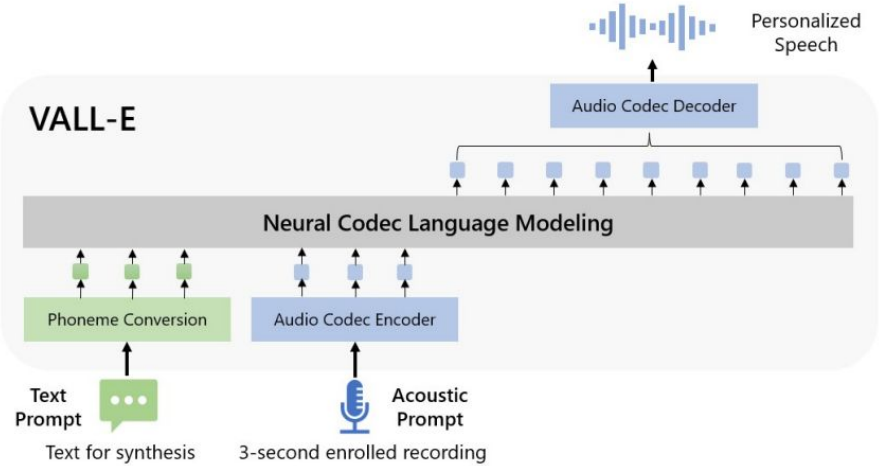


Figure 1: The overview of VALL-E. Unlike the previous pipeline (e.g., phoneme → mel-spectrogram → waveform), the pipeline of VALL-E is phoneme → discrete code → waveform. VALL-E generates the discrete audio codec codes based on phoneme and acoustic code prompts, corresponding to the target content and the speaker's voice. VALL-E directly enables various speech synthesis applications, such as zero-shot TTS, speech editing, and content creation combined with other generative AI models like GPT-3 [Brown et al., 2020].

*These authors contributed equally to this work. Correspondence: {yuwu1,shujliu,fuwei}@microsoft.com

Решаемая задача и вклад в отрасль

- Задача: **Text to Speech** (TTS)
- Новизна: воспроизведение **голоса из** короткого **входного промпта**

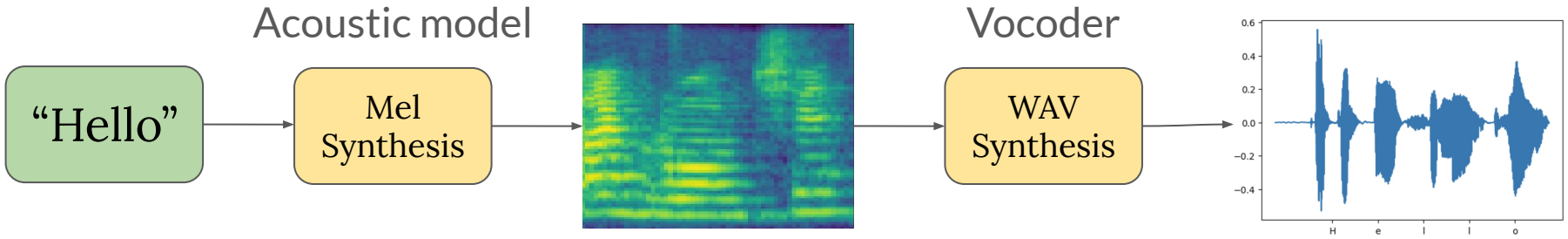


Авторы

Chengyi Wang	Sanyuan Chen	Yu Wu
❏ Speech recognition ❏ Speech translation	❏ Pre-training ❏ Speech ❏ NLP	❏ Speech Recognition ❏ Conversational AI ❏ Pre-Training

Text to Speech

- WaveNet (2016, vocoder)
- Tacatron (2017), Tacatron 2 (2018)
- FastSpeech (2019), FastSpeech 2 (2022)

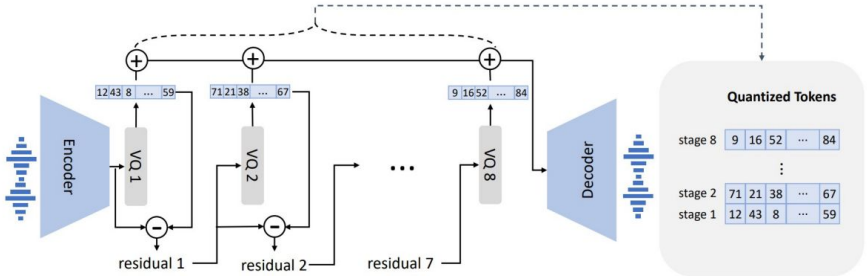


Zero-Shot TTS

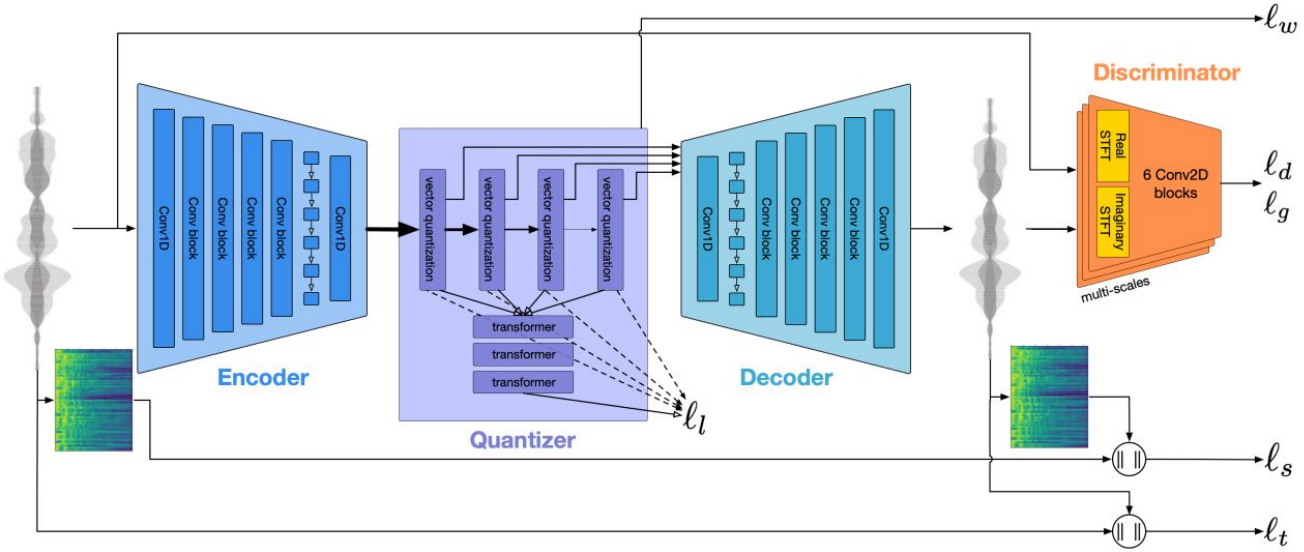
- Speaker adaptation
  - Chen et al. [2019], Wang et al. [2020], Chen et al. [2021]
- Speaker encoding
  - Jia et al. [2018], Arik et al. [2018], Wu et al. [2022]

VALL-E

- Обучение на огромном и разнообразном датасете
- Обычный пайплайн TTS но с признаками из кодек-модели вместо мел-спектрограммы в качестве промежуточного представления
- Предобученные акустическая модель (EnCodec) и вокодер (DeCodec)



EnCodec



[High Fidelity Neural Audio Compression]

Работы из области

- AudioLM [2022] (audio codecs, speech-to-speech)
- VQTTS [2022] (audio codecs, один голос)
- GradStyleSpeech [2022] (diffusion, вокодер: HiFiGAN)
- YourTTS [2022] (прошлая SoTA, брали как бейзлайн, вокодер: HiFiGAN)

Цитирования

Всего 5 цитирований:

- VALL-E X [2023] (**продолжение**)
- InstructTTS [2023] (**конкуренты**)
- SPEAR-TTS [2023] (**конкуренты**)
- An AI Grand Challenge for Education [2023] (VALL-E как пример голоса для ИИ-учителя)
- Necrorobotics. The Ethics of Personalised Resurrection [2023] (про этическую сторону использования VALL-E (ИИ в целом))

Рецензия

Плюсы:

- Наличие удобного **демо** с результатами
- **Понятное** и **последовательное** изложение материала
- Наличие сравнения с прошлой SoTA моделью на разных наборах данных

Минусы:

- Относительно **долгий инференс** из-за авторегрессионной части
- Хотелось бы увидеть сравнение с работами, вышедшими примерно в то же время