Обзор XTTS-v2

Описание технологии

XTTS-v2 — это модель для генерации голоса, основанная на последних разработках в области генеративного искусственного интеллекта. Она обеспечивает очень выразительную озвучку, более качественное клонирование голоса с помощью короткого аудиоклипа и все расширенные функции Coqui Studio.

XTTS-v2 вдохновлёна большими языковыми моделями и создана специально для разработчиков игр. Однако данная технология ориентирована на исключительную производительность преобразования текста в речь.

Примеры использования

- Сервисы электорнных и аудиокниг — для создания аудиокниг, подкастов и обучающих материалов.
- Переводчики и пользователи видеохостингов — для создания и просмотра иноязычного контента с помощью синтезированного голоса на интересующем языке.
- Разработчики игр — для создания персонажей с реалистичным голосом и диалогами.

Отзывы в интернете

“У XTTS очень качественная генерация голоса. Если вы повторяете предложение, оно каждый раз звучит по-разному, а если вы объединяете его с вашей голосовой моделью RVC, результат будет просто великолепным. Будущее должно быть с открытым исходным кодом!”

“Удачи в аду зависимостей в попытках установить это”

“Прощайте, ElevenLabs. Эта модель XTTS потрясающая. Спасибо!”

Тестовые примеры

1. Клонирование голоса для озвучивания заданного текста.

- Модели было необходимо воспроизвести голос Сергея Бурунова и озвучить фразу:
  «Твои мысли подобны кругам на воде, друг мой. В волнении исчезает ясность, но если ты дашь волнам успокоиться, ответ станет очевидным.»
  Пример голоса для клонирования:

Пример голоса для клонирования

- Для сравнения приведены 3 полученных аудиофайла. Как и говорится в документации к

XTTS, для одних и тех же данных модель выдаёт разные варианты речи.

Результат генерации 3
Результат генерации 2
Результат генерации 1

---

2. Перевод текста с сохранением интонации:

- Модели было необходимо воспроизвести всё тот же голос Сергея Бурунова, но озвучить фразу уже на английском языке. Текст для модели переводился вручную:
«The warrior's fame as the best of the best, of the best, of the best followed him closely. Stop talking, let's fight! And even after seeing the legendary sword of heroes,
our warrior did not blink an eye. He was such a great master!»

Результат перевода

---

3. Замена голоса в аудио:

- Модели было необходимо заменить голос Сергея Бурунова на голос Совы из мультфильма про Винни-Пуха, сохранив при этом оригинальную интонацию. Входная аудиозапись для модели та же, что и в предыдущих пунктах.
Текст для модели: «Слава воина как лучшего из лучших, из лучших, из лучших преследовала его по пятам. Хватит болтать, давай сразимся! И даже узрев легендарный меч героев, наш воин не моргнул и глазом. Такой он был, великий мастер.»

Пример голоса для клонирования
Результат замены голоса

Возможность автономного запуска

Модель XTTS-v2 использует диффузионные модели нейронных сетей для перевода выхода GPT-модели в фрейм спектрограммы и модель UnivNet для генерации окончательного аудиосигнала.
Поэтому без изменения архитектуры не получится использовать нейросеть Mistarl NeMo для обеспечения функциональности, предложенной XTTS-v2.

Модель распространения

Полная бесплатная версия XTTS-v2 представлена в виде репозитория с исходным кодом на GitHub и не ограничивает пользователя в наборе инструментов.
Если нет возможности установить и использовать продукт локально, можно воспользоваться его демо-версией — Ссылка на демо-версию XTTS-v2.
В таком случае пользователю доступен полный набор функций модели, но генерируемый текст

должен быть не длиннее 200 символов.

Преимущества XTTS-v2

- Широкий выбор языков для распознавания и клонировния, в том числе и русский.
- Для работы достаточно короткого фрагмента аудиозаписи - от 6 секунд.
- Доступна демо-версия.
- Есть возможность клонировать интонацию и стиль речи вместе с голосом.
- Межъязыковое клонирование голоса.

Недостатки XTTS-v2

- Нет возможности самостоятельно задавать интонацию речи.
- Необходимость подбирать подходящую модель, обученную на интересующем языке, чтобы при генерации голоса не было акцента.
- Есть погрешности в произношении, интонации, ударениях.
- Голоса не всегда похожи после клонирования.
- Обработанные голоса персонажей фильмов и мультфильмов модель клонирует, но пытается приблизить к человеческим.

Последняя дата обновления

Последняя дата обновления на GitHub: 12.12.2023