Отчет по оценке качества синтеза моделью **parler-tts-mini-jenny-30H**

Пушин Максим, ВШЭ, Прикладная Математика и Информатика, 3 курс

Метрики для оценки качества синтеза речи делятся на субъективные и объективные, и оценивают модели основном первыми. Но в моем же отчете я старался работал с обоими в равной степени. В качестве объективной метрики я использовал WEB – метрика характеризующая надежность, и показывающая коэффициент ошибок. Также я обращал внимание на качество речи, для оценки которой обычно используются опросы, и ее зависимость от описания.

Мой план состоял в том, чтобы сначала разобраться в работе модели, а именно сделать первый синтез с данными из примера, далее смотреть как будет меняться качество речи при изменении описания. Следующим шагом я планировал проверить надежность, дать модели много текстов, не особо думая о описании, чтобы проверить насколько корректно она синтезирует речь. Последним шагом я планировал взять реальную запись, и попробовать подобрать описание так, чтобы модель выдала наиболее похожий результат.

Ход работы:

1. Проделав установку и создание первой записи, я пробовал менять описание и смотрел, как меняется качество речи. Я пробовал менять скорость речи, акценты, тембры. В первую очередь я попробовал менять скорость речи, и это модель отлично исполнила. Однако изменений в эмоциональной составляющей я не заметил. Далее я пробовал менять тембр голоса и акценты, и в это аспекте мне изменения виделись уже существенней.

2. Вторым шагом была проверка надежности. Для создание скриптов, я обратился к ChatGPT и он мне сгенерировал 10 случайных текстов с описаниями. Сами тексы лежат в prompts.txt. Далее я генерировал речь, и после с помощью библиотеки SpeechRecognition делал ее транскрипцию. После убирал все знаки препинания и переводил в один регистр, получал коэффициент ошибок с помощью библиотеки jiwer. Для уменьшения погрешности, в случае неправильной работы транскриптора, например, я повторял цикл несколько раз и брал среднее.
   Результат: был получен WER равный 0.13, что является далеко не идеальным результатом. Однако, в процессе живого прослушивания я обнаружил, что иногда, при отсутствии точного описания или при его непонятности, модель может пропускать целые фрагменты данного ей текста. Но при запуске на полноценно воспроизведенных примерах, коэффициент ошибки почти всегда был < 5%, и проблемы возникали только с воспроизведением только первого и/или последнего слова.

3. Последним шагом было попробовать подобрать описание так, чтобы сгенерированная речь была похожа на оригинал по темпу и паузам. В качестве примера речи использовался фрагмент аудиокниги The Queen of Nothing by Holly Black. С надежностью снова возникала проблема пропуска последнего слова, однако в целом текст передавался правильно(оценивал без транскриптора, вживую). Аудио автора – real_actor_1.wav.
   Сначала я попробовал синтезировать речь без какого-либо описания и получил вариант parler1.wav. Модель читала быстро, без каких-либо смысловых остановок.

Следующим шагом я описывал общие характеристики оригинала: скорость, наличие пауз. Тогда у меня получился вариант parler2.wav, который уже очень был похож на оригинал (не считая пола диктора, конечно).

Далее я указывал модели, где конкретнее делать паузы или на каких словах акцентировать ударение, однако по ощущением именно вариант 2 дал наибольшую точность.

Код, а также сгенерированные аудио лежат в github репозитории.

Вывод: Модель правдоподобно и с высокой надежностью способна синтезировать речь. Однако в процессе работы с ней я встречал некоторые проблемы. Среди проблем с надежностью, выделю пропуск последнего и реже первого слова при синтезе речи. Так не могу не отметить единичные проблемы, когда модель выдает только аудио из 2-3 несвязных слов, что плохо влияет на автономность. С точки зрения качества речи, как я уже говорил, она достаточно правдоподобна. Модель хорошо справляется с изменением скорости речи, ее тембром и наличием пауз в тексте. Однако я не заметил существенных различий в интонации синтезируемой речи и акцентах, на указанных словах в тексте.