УДК 004.89

Вдовиченко М. С.
Научный руководитель – канд. физ.-мат. наук Азимов Р. Ш.
Университет ИТМО
wdowitschenko.mischa@yandex.ru

Работа выполнена в рамках темы НИР «Разработка мультиязычной модели синтеза речи на малоресурсных языках».

Введение

Модели для синтеза речи представляют собой каскад последовательно применяемых моделей [1,2]. Для обучения современных моделей, обладающим достаточно большим числом параметров порядка сотни миллионов, необходим большой объем данных. Однако не всегда удается собрать датасеты достаточного объема, при обучении мультиязычных моделей зачастую приходится иметь дело с дисбалансом данных по используемым языкам. Использование схожих языков, причисленных к одной языковой группе, при обучении модели может повысить качество синтезируемого звука на малоресурсном языке. Также для решения проблемы дисбаланса данных можно, например, или размножить малоресурсные языки или использовать взвешенную функцию потерь.

Для задачи синтеза речи обучающие данные в простейшем виде представляют собой набор пар, состоящий из аудиозаписи, соответствующий ей текст и метки, определяющей язык. Для сбора данных могут применяться открытые источники, так и закрытые, например, записи одного диктора - такие записи обладают лучшим качеством и используются в финальных стадиях обучения модели.

В рамках данной работы решается задача построение пайплайна для турецкого и узбекского языков на проприетарных датасетах и создание мультиязычной модели синтеза речи на основе архитектуры tortoise [3]. Проводились эксперименты по подбору входного набора данных, оптимальных этапов обучения и числа шагов для финального обучения для получения наилучшего качества претрейна модели синтеза речи в смысле объективных и субъективных метрик.

Основная часть

Производилось обучение wav-токенайзера на турецких и турко-узбекских данных. Замерены метрики при попытке восстановления звука, произведено сравнение чекпоинтов моделей между собой на каждом из языков и сравнение с референсными аудио-дорожками, при использовании их для подсчета метрик.

В качестве обучаемой модели бралась архитектура тортойза - для генерации используется авторегрессионный трансформер llama архитектура. Сначала проводились эксперименты по подбору оптимального датасета для получения претрейна, для сравнимости экспериментов фиксировалось количество эпох обучения претрейна, равное 3, при обучении на спикера фиксировалось количество эпох как 10.

Было проверено примерная равнозначность клонирования данных и использование соответствующего коэффициента в взвешенной функции потерь, были рассмотрены различные конфигурации претрейнов с различными коэффициентами в в взвешенной функции потерь, в дальнейшем запись n к m следует трактовать как множитель n в функции потерь для турецких данных и множитель m для узбекских данных. Рассматривались варианты коэффициентов 1 к 1, 1 к 4, 1 к 5, 1 к 10.

Также рассматривается оптимальность обучения модели в смысле лосса финальной валидационной корзины от данных спикера на получаемое финальное

качество модели. Рассматривается также возможные стадии промежуточного обучения в целях повышения финального качества модели для синтеза речи. Качество получаемых моделей замеряется на специально подготовленных целевых корзинках на соответствующих языках, при этом для анализа используются такие метрики как WER, SIMilarity, AES [4], UTMOS и PSER.

Выводы

Выявлено, что добавление турецких данных позволяет улучшить качество получаемой речи и звука на узбекском языке. Найдено лучшее соотношение датасетов по получаемому финальному качеству модели на узбекском языке, качество получаемой модели сначала увеличивается при увеличении доли узбекских данных в итоговом датасете, затем снижается. Это объясняется низким качеством звука в узбекских данных и достаточно хорошим качеством звука в турецких данных, постепенное увеличение доли узбекских данных сначала улучшает качество синтезируемой речи на узбекском языке, при этом качество звука остается хорошим, при дальнейшем увеличением доли узбекских данных наблюдается деградация получаемой модели, по качеству звука и соответственно качеству речи. Лучшей конфигурацией обучения с точки зрения претрейна является 1 к 5, удается достичь значения по PSER модели, используемой для Алисы на узбекском языке, а именно доля записей ошибочно озвученных составляет не более 20 процентов. Более длительное обучение показывает лучшие результаты по метрикам как для совместного претрейна 1 к 5, так и для претрейна при использовании только узбекского языка. Промежуточные стадии обучения не показывают ожидаемого существенного улучшения модели, из-за достижения предела вместимости модели. На турецком языке по метрике PSER наблюдается улучшение качества модели при использовании совместного претрейна, однако достичь качества продовой модели на турецком языке (модель используется в турецком навигаторе) пока не получилось, а именно моя модель показывает PSER 0.18, используемая на данный момент модель показывает 0.14.

Литература

1. End-to-End Text-to-Speech (TTS) [Электронный ресурс]. – Режим доступа: https://www.arunbaby.com/speech-tech/0039-end-to-end-tts/ (Дата обращения 26.02.2026).
2. Виртуальный рассказчик 2.0: эволюция нейросетевого рассказчика в Яндекс Книгах [Электронный ресурс]. – Режим доступа: https://habr.com/ru/amp/publications/936250/ (Дата обращения 26.02.2026).
3. Better speech synthesis through scaling [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2305.07243 (Дата обращения 26.02.2026).
4. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2502.05139 (Дата обращения 26.02.2026).

«26» февраля 2026 года _____ Вдовиченко М.С.

Подпись автора расшифровка подписи

_____ Азимов Р.Ш.

Подпись научного руководителя расшифровка подписи