Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Анализ моделей нейронных сетей для распознавания речи

Студент ИУ5-34М _____ **Старых Ф.А.**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Ю.Е. Гапанюк**
(Подпись, дата) (И.О.Фамилия)

2023 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ___ИУ5____
(Индекс)
_____ ___В.И. Терехов__
(И.О.Фамилия)
«_04_» ____сентября_____ 2023 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме ___Анализ моделей нейронных сетей для распознавания речи_____

Студент группы _ИУ5-34М__

_____Старых Фёдор Артемович_ _____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
_____ИССЛЕДОВАТЕЛЬСКАЯ_____
Источник тематики (кафедра, предприятие, НИР) _____КАФЕДРА_____

График выполнения НИР: 25% к ____ нед., 50% к ____ нед., 75% к ___ нед., 100% к ____ нед.

Техническое задание___Исследовать и провести анализ существующих моделей нейронных сетей для распознавания рячи_____
_____
_____
_____
_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __12___ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
_____
_____
_____

Дата выдачи задания «_04_» ____сентября_____ 2023 г.

Руководитель НИР _____ _____Ю.Е. Гапанюк__
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____Ф.А. Старых_____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Введение

Технология Speech-to-text(STT) отвечает за перевод голоса в текст с помощью нейросети. В основе лежит многоуровневый процесс обработки и анализа аудиосодержимого. Речь с помощью искусственного интеллекта преобразуется в буквы, слова, фразы и предложения, и на выходе получается текстовая версия аудио.

В основе работы технологии STT — нейросети, которые обрабатывают речь и возвращают распознанный текст. Речь состоит из звуков, а текст состоит из букв. Основная задача нейросети — распознать, какой букве соответствует рисунок на спектрограмме аудиозаписи, затем преобразовать отдельные буквы в слова, а слова — в полноценные предложения.

Чтобы научиться распознавать среди звуков буквы, инженеры обучают нейросеть на подготовленном датасете. Датасет состоит из аудиозаписей с голосом, которые сопровождаются размеченным текстом. Таким образом, на вход нейросети подаётся пара аудио-текст, из которой она должна найти соответствие «рисунку» аудиодорожки определенных букв и слов.

В процессе обучения искусственный интеллект разбивает запись с голосом на короткие отрезки и пытается предсказать по спектрограмме каждой из них, что это за буква. При этом в процессе предсказания нейросеть не выдаёт однозначный результат: она определяет, с какой вероятностью перед ней та или иная буква.

Когда вероятности по каждой букве в записи голоса вычислены, искусственный интеллект пытается понять, какое это слово. Для этого есть контекст — или, проще говоря, словарь, — с которым нейросеть проводит сравнение вероятных букв. В результате получается набор распознанных слов.

Слова, в свою очередь, искусственный интеллект складывает в предложения. Финальный этап — это смысловая обработка. Кроме

непосредственно распознавания, важно, чтобы текст на выходе был связным, осмысленным и правильно оформленным (был поделён на предложения, имел знаки препинания).

Связность и осмысленность в технологии распознавания речи обеспечивается, в том числе, объёмом текстов, которые нейросеть обработала на этапе обучения. Например, если в момент распознавания близки вероятности слов «еду» и «иду», то при построении полной фразы «я еду на машине» нейросеть выберет верный вариант, потому что слова «еду» и «машина» ближе по контексту, чем «иду» и «машина».

Существующие модели нейронных сетей распознавания речи

В наше время существует множество различных моделей нейронных сетей для распознавания речи. Для сравнения возьмем некоторые из них. Такие модели применяются во множестве сфер: бизнес, различные сервисы и платформы, чат-боты, голосовое управление и многие другие.

Silero

Сейчас для всех желающих доступны два сервиса для распознавания речи:

- Телеграм-бот для коротких и не очень длинных аудио;
- Сервис audio-v-text.silero.ai для более длинных аудио, в котором можно скачать отчет в виде exel-таблицы.

Сервис написан и работает на собственном движке распознавания речи, без проксирования во внешние сервисы и с минимально возможным количеством зависимостей. В случае нарушения связности возможен оперативный перевод хостинга в другие регионы.

Sber(Salute Speech)

- В приложении есть два раздела. «Распознавание» — для текстовой расшифровки голосовых файлов. «Синтез» — для озвучивания текста с возможностью настраивать паузы и ударения. Синтезировать текст можно

разными голосами из семи вариантов на русском и английском языках. Сервис распознаёт аудио в шести форматах: pcm, opus, mp3, flac, alaw, mulaw.

•В приложение встроен GigaChat API, поэтому пользователь может загрузить короткие тезисы, нейросеть по ним подготовит текст для озвучивания. Также с помощью GigaChat в приложении можно сделать короткую выжимку длинного текста, а после озвучить материал.

Google

Google Cloud Speech-to-Text — это продвинутый инструмент для автоматического преобразования речи в текст и транскрипции. Это полезный сервис, который позволяет разработчикам использовать автоответчики в колл-центрах, позволяет IoT-устройствам общаться с пользователями и преобразовывать текстовые сообщения в голосовой формат.

Speech-to-Text, ранее называвшийся Cloud Speech API, был впервые выпущен в 2016 году. По данным Google, в первые годы его работы, использование API удваивалось каждые шесть месяцев. Это решение основано на самых передовых алгоритмах нейронной сети глубокого обучения Google для автоматического распознавания речи (ASR).

Есть возможность быстро развернуть ASR в облаке с помощью API или даже локально с помощью локального преобразование речи в текст, которое интегрирует технологии распознавания речи Google в ваше локальное решение. В ответ на необходимые правила к размещению данных и соответствию требованиям, вы можете взять под контроль свою инфраструктуру, одновременно извлекая выгоду из технологии распознавания речи с высокозащищенными речевыми данными.

Yandex

SpeechKit предоставляет два способа распознавания речи:

- Потоковое распознавание применяется для распознавания в режиме реального времени. При потоковом распознавании SpeechKit получает короткие аудиофрагменты и отправляет результаты, в том числе промежуточные, в рамках одного соединения.

- Распознавание аудиофайлов. SpeechKit может распознавать аудиозаписи в синхронном и асинхронном режиме.

  - Синхронное распознавание имеет жесткие ограничения на размер и длительность файла и подходит для распознавания одноканальных аудио до 30 секунд.
  - Асинхронное распознавание может обрабатывать многоканальные аудиозаписи. Максимальная длительность файла — 4 часа.

Tinkoff

Речевые технологии Tinkoff VoiceKit — это глубокие нейросетевые модели для синтеза и распознавания речи, которые в течение последних лет разрабатывались в Тинькофф в рамках стратегии AI First.

Технология Tinkof VoiceKit может использоваться, например, для:

- Создания собственных голосовых помощников
- Создания роботов для автоматизации работы колл-центра
- Быстрой записи аудиокниг, озвучки и редактирования видеороликов
- Построения системы речевой аналитики по транскрибированным текстам — например, в колл-центрах для контроля работы операторов
- Создания приложений для людей с ограниченными возможностями
- Транскрибирования любых звуковых записей публичных выступлений

•Поисковой оптимизации и полнотекстовому поиску по аудио и видеозаписям

Все модели - это модели упакованные в production сервисы. Используемый показатель — WER (для простоты восприятия можно мысленно пририсовать знак процента или считать WER процентом ошибок в словах). Сравнение проведено на основе различных вариантов датасетов, тематик и, соответственно областей применения.

Таблица 1. Сухие метрики

Датасет	Google	Sber	Silero	Tinkoff	Yandex
Чтение	11	7	7	8	13
Умная колонка	24	6	30	27	14
Энергосбыт	39	20	16	15	13
Звонки (такси)	16	32	13	21	15
Публичные выступления	27	18	14	20	21
Финансы (оператор)	37	33	25	23	22
Аэропорт	36	26	21	25	21
Аудио книги	60	19	24	28	22
Радио	61	26	18	27	23
Умная колонка (далеко)	49	8	41	52	18
Банк	30	28	39	28	25
Заседания суда	29	31	20	31	29
Финансы (клиент)	55	67	38	33	32
YouTube	50	34	28	38	32
Медицинские термины	37	50	35	42	38

Диспетчерская	68	54	41	43	42
Стихи, песни и рэп	70	61	43	56	54
Справочная	50	32	25	27	-

Как можно заметить, что каждый силен в том домене, на котором фокусируется. Tinkoff — на звонках в банк, справочную, финансовые сервисы. Сбер имеет ультимативно лучшие результаты на своей "умной колонке" и в среднем имеют неплохие показатели. Модель Сбера на доменах, где оригинальные данные находятся в диапазоне 8 kHz, показывает себя достойно, но она не ультимативно лучшая. Яндекс с недавнего времени сделал значительный прогресс и показал более лучшие результаты чем ранее. В данном сравнении Google является аутсайдером и в среднем показал не самые лучшие и по большей части наихудшие результаты. Лидером данного сравнения оказалась модель Silero

Проведем небольшой сравнительный анализ на основе полученных результатов. Посчитаем количество доменов, где модели поставщика лучшие / худшие (допускается "послабление" в 10% от лучшего или худшего результата):

Таблица 2. Сравнительный анализ лучше/хуже всех

Сервис	Лучше всех	Хуже всех
Google	1	11
Sber	3	4
Silero	12	3
Tinkoff	6	1
Yandex	5	2

Как и ожидалось — Silero показывает в среднем неплохие показатели на всех доменах, заметно отставая на банках и финансах. Также если смотреть по

формальной метрике "на каком числе доменов модель лучшая или почти лучшая" — то данная модель как минимум лучше всех генерализуется. Неудачи модели вытекают из отсутствия личной умной колонки, а так же отсутствием доступа к банковским данным.

Заключение

В ходе данного исследования были проанализированы и сравнены некоторые популярные в РФ модели нейросетей для распознавания речи. Исследование показало, что относительным лидером можно считать модель Silero, в то время как худшие показатели были у модели Google. Безоговорочно стоит учитывать тот факт, что каждая модель разрабатывалась для определенных целей и по большей части в области своей специализации каждая из оцененных моделей показала лучший или хороший результат.

Тенденция создания, совершенствования и разработки моделей распознавания речи в наше время пользуется спросом и имеет довольно большую популярность. Практически все модели пытаются стремиться к совершенству как минимум в сфере, для которой та или иная модель предназначается.

В заключение, разработка моделей нейронных сетей для распознавания голоса является актуальной и перспективной задачей в современной информационной технологии.

Проанализированные модели нейронных сетей имеют широкий спектр применений, начиная от голосовых помощников и умных домашних устройств, заканчивая системами безопасности и медицинскими приложениями. Они могут быть использованы для автоматизации повседневных задач, улучшения пользовательского опыта и повышения эффективности работы в различных сферах.

Однако, несмотря на достигнутые результаты, существует ряд вызовов и проблем, которые требуют дальнейших исследований и разработок. Например, улучшение моделей для работы с различными акцентами и диалектами, а также повышение устойчивости к шуму и фоновым звукам.

В целом, разработка моделей нейронных сетей для распознавания голоса имеет большой потенциал и может существенно улучшить нашу повседневную жизнь.

Список использованных источников

1. https://silero.ai/
2. https://www.tinkoff.ru/software/voicekit/
3. https://vc.ru/services/917940-sber-vypustil-prilozhenie-salutespeech-app-dlya-sinteza-i-raspoznavaniya-rechi-v-audio
4. https://cloud.yandex.ru/ru/docs/speechkit/stt/
5. https://developers.sber.ru/help/salutespeech/how-speech-recognition-works
6. https://cloud.ru/ru/aicloud/salutespeech