СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ РАСПОЗНАВАНИЯ РУССКОЯЗЫЧНОЙ РЕЧИ НА ПРИМЕРЕ ТЕЛЕФОННЫХ ЗВОНКОВ

Газизулин Р.М., магистрант
Хартьян Д.Ю., к.тех.н., доцент,
ТюмГУ, г. Тюмень, Россия

Аннотация. Рассмотрены модели Wav2Vec2-Large-Ru-Golos, Wav2Vec2-Large-Ru-Golos-With-LM, Whisper-Turbo, GigaAM-RNN, GigaAM-CTC для распознавания русскоязычной речи. Приведены результаты по точности, скорости и устойчивости к шуму. Выявлены сильные и слабые стороны каждой модели.

Ключевые слова: транскрибация, WER, телефонные звонки.

Введение. Автоматическое распознавание речи (АРР) является ключевым инструментом в современной обработке речевых данных, с широким применением в сферах обслуживания клиентов, управления интеллектуальными устройствами и других бизнес-процессах.В статье "Исследование возможностей и оценка качества распознавания звучащей речи различными нейросетевыми моделями" был проведен анализ производительности ряда нейросетевых моделей на стандартных датасетах. Авторы сделали акцент на сравнении архитектур и подходов, определяя области их эффективного применения. Ключевыми выводами стали выявление преимуществ некоторых моделей для специфических условий распознавания и рекомендации по выбору моделей в зависимости от задач [2]. Отличительной особенностью представленной статьи является анализ моделей АРР на собственных данных, что делает результаты более релевантными для реальных бизнес-процессов, где качество распознавания речи напрямую связано с особенностями речевой коммуникации в телефонных разговорах. Таким образом, в рамках данного исследования проводится:

1. Сравнительный анализ моделей на телефонных звонках — особом формате речи, отличающемся качеством записи и спецификой диалогов.

2. Оценка устойчивости моделей к шуму, что является важным фактором для применения в реальных условиях, особенно в контактных центрах и системах голосового управления.

Сравнение архитектур. Архитектура модели Wav2Vec2-Large-Ru-Golos основана на трансформерах с предобучением на больших объемах данных без разметки. Модель обучена на корпусе русскоязычной речи. Основное преимущество — использование контекстных представлений звуковых сигналов для точного распознавания. Модель содержит 300 миллионов параметров.Существует модифицированная версия Wav2Vec2-Large-Ru-Golos с подключением языковой модели. LM улучшает точность распознавания, особенно для длинных текстов, но увеличивает время обработки и ресурсоемкость. Whisper large-v3-turbo — это оптимизированная версия модели Whisper large-v3 с уменьшенным количеством декодирующих слоёв [4]. Архитектура основана на трансформерах с мультиязыковой поддержкой, что делает модель универсальной для различных сценариев. Модель содержит 809 миллионов параметров. GigaAM (GigaAcoustic Model) представляет собой базовую акустическую модель, которая использует Conformer-энкодер с приблизительным количеством параметров в 240 миллионов. Версия GigaAM-CTC была адаптирована для обучения с применением CTC-функции потерь и посимвольной токенизации, тогда как GigaAM-RNNT прошла дообучение, используя RNN-T-функцию потерь и токенизацию на уровне подслов [3].

Оценка моделей. Оценка моделей производилась по трем ключевым параметрам: точность (WER), быстрота работы и устойчивость к шуму.Условия тестирования: объем данных составил 1000 аудиофайлов телефонных разговоров длительностью от 10 до 456 секунд, тестирование проводилось на GPU Tesla K80.Предобработка аудиоданных включала нормализацию громкости для устранения различий в уровне сигнала, базовую фильтрацию шума (кроме тестов на устойчивость к шуму, где добавлялся искусственный

шум), а также приведение всех аудиофайлов к монофоническому формату с частотой дискретизации 16 000 Гц.Ключевым показателем точности является WER (Word ErrorRate) [5]. Результаты (таблица 1) приведены как для записей без шума, так и для зашумленных (SNR 10 дБ и 5 дБ). SNR – это величина, равная отношению сигнала и шума [1]. Представлен график сравнения точностей моделей (рис. 1).

Таблица 1 – Показатель ошибкимоделей в различных условиях шума, %

Модель	WER (без шума)	WER (с SNR 10 дБ)	WER (с SNR 5 дБ)
Wav2Vec2-Large-Ru-Golos	74.20	-	-
Wav2Vec2-Large-Ru-Golos-With-LM	77.21	-	-
Whisper-Turbo	42.05	55.98	61.59
GigaAM-RNN	30.96	45.15	50.03
GigaAM-CTC	32.80	48.00	54.84

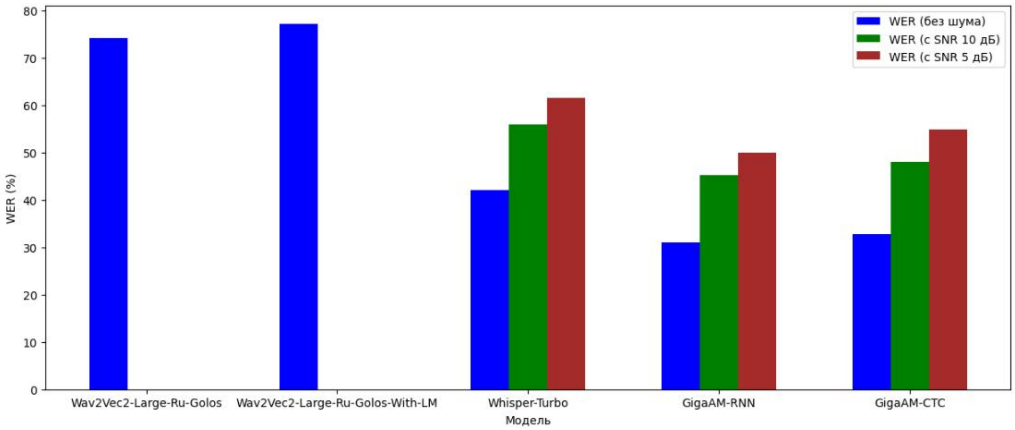

Рисунок 1. Сравнение WERмоделей в различных условиях

Wav2Vec2 модели оказались непригодны для обработки зашумленных данных, даже при использовании языковой модели (LM). GigaAM-RNN и GigaAM-CTC показали высокую устойчивость, сохраняя разумную точность. Whisper-Turbo продемонстрировал хорошую устойчивость, но снизил точность при шуме.Длина аудиофайлов не влияет на WER (рис.2), так как перед обработкой файлы сегментируются на отрезки фиксированной длины. Сегментация устраняет зависимость точности от длины записи, так как каждый сегмент обрабатывается отдельно. Транскрипция выполняется частями, что обеспечивает стабильность результата.

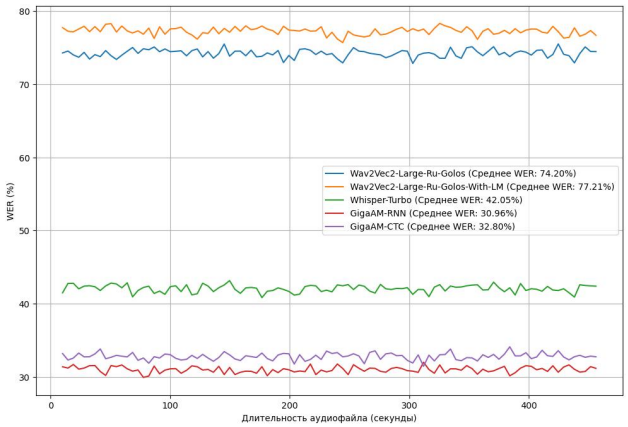

Рисунок 2. Зависимость WER от длины аудиофайлов для различных моделей

Небольшие колебания WER, наблюдаемые на графике, могут быть объяснены статистическими погрешностями, связанными с вариациями в данных и случайными ошибками моделей. Колебания не указывают на систематическую зависимость WER от длины аудиофайлов, а скорее отражают естественные флуктуации. В исследовании представлено сравнение среднего времени обработки для различных моделей (таблица 2). Параметр времени обработки оценивается как среднее время на 1 минуту аудио. Представлен график сравнения точностей моделей (рис. 3).

Таблица 2 – Среднее временя обработки аудио моделями

Модель	**Среднее время обработки**
Wav2Vec2-Large-Ru-Golos	~63 секунды
Wav2Vec2-Large-Ru-Golos-With-LM	~61 секунда
Whisper-Turbo	~3 секунды
GigaAM-RNN	~18 секунд
GigaAM-CTC	~20 секунд

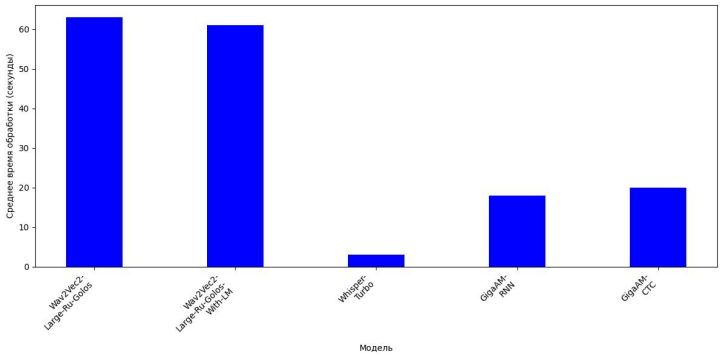

Рисунок 3. Скорость обработки аудио моделями

Whisper-Turbo значительно превосходит другие модели по скорости, обеспечивая почти реальное время обработки. GigaAM модели обеспечивают приемлемую производительность. Wav2Vec2 модели требуют больше времени на обработку.

Выводы. Модели GigaAM показали лучший баланс точности и устойчивости к шуму. GigaAM-RNN обеспечивает превосходную устойчивость к шуму благодаря архитектуре, ориентированной на последовательную обработку данных, и делает ее лидером по точности как в чистых, так и в зашумленных данных. Более длительное время обработки, в сравнение с Whisper, связано с тем, что RNN обрабатывает данные последовательно, что требует больше вычислительных ресурсов. Метод CTC оптимизирован для точной временной привязки, что обеспечивает сбалансированный вариант, показывая приемлемую точность.Высокий показатель WER Wav2Vec2 моделей указывает на слабую адаптацию моделей к тестовым данным, что может быть связано с недостаточным объемом данных для обучения или несовершенной языковой моделью. Wav2Vec2 основаны на трансформерах с предварительным обучением и показывают низкую точность даже без шума на наших аудиоданных, что сделало их тестирование в зашумленных условиях нецелесообразным. Whisper-Turbo использует мультиязыковую архитектуру трансформеров, что обеспечивает высокую скорость и универсальность. Одной из ключевых причин, почему трансформеры быстрее рекуррентных нейронных сетей, является их способность к параллельным вычислениям. Однако снижение точности в условиях шума указывает на меньшую устойчивость к зашумленным данным, возможно, из-за ограниченной обработки шумовых сценариев в обучении или недостаточной локальной оптимизации.

Рекомендации. Для проектов, требующих высокой точности даже в шумной среде, использовать GigaAM-RNN. Для систем реального времени с большими объемами данных подойдет Whisper-Turbo. Для универсальных задач со средними требованиями можно рекомендовать GigaAM-CTC. Если

необходимо расширить исследование, например, включив тесты на других языках или условиях записи, рекомендуется масштабировать анализ с дополнительными параметрами.

Литература

1. Отношение сигнал/шум — Рувики: Интернет-энциклопедия. – URL: https://ru.ruwiki.ru/wiki/Отношение_сигнал/шум (дата обращения: 27.11.2024).

2. Ушакова, А. А. Исследование возможностей и оценка качества распознавания звучащей речи различными нейросетевыми моделями / А. А. Ушакова, В. В. Гаршина // Труды молодых учёных факультета компьютерных наук ВГУ : Сборник статей. – Воронеж : Воронежский государственный университет, 2024. – С. 540-546. – EDN IKLEWS.

3. GigaAM: класс открытых моделей для обработки звучащей речи / Хабр. – URL: https://habr.com/ru/companies/sberdevices/articles/805569/ (дата обращения 27.11.2024).

4. Openai/whisper-large-v3-turbo · Hugging Face. – Электронный ресурс. – URL: https://huggingface.co/openai/whisper-large-v3-turbo (дата обращения 27.11.2024).

5. Shentsov, Ya. A. Application of speech recognition and autoreference models for logging tasks / Ya. A. Shentsov, T. Y. Chernysheva, G. B. Barskaya // Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023), Dushanbe, 20–22 декабря 2023 года. Vol. 13065. – Washington: SPIE-SOC PHOTO-OPTICAL INSTRUMENTATION ENGINEERS, 2024. – P. 1306503. – DOI 10.1117/12.3024859. – EDN WDHKRT.