М. А. Ефимова, И. С. Герасименко, М. С. Воробьева
ФГАОУ ВО «Тюменский государственный университет», г. Тюмень
УДК 004.9

РАЗРАБОТКА СЕРВИСА ПРЕОБРАЗОВАНИЯ РУССКОЯЗЫЧНОЙ РЕЧИ В ТЕКСТ: ТЕХНОЛОГИИ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ (ASR)

Аннотация. В статье предложено контейнеризированное On-Premise решение для преобразования русскоязычной речи в текст с REST-интерфейсом на базе модели GigaAM-RNNT-v2. Эксперимент на 110 студенческих голосовых сообщениях показал высокий уровень точности (низкий WER/CER) без облачных сервисов.

Ключевые слова: обработка звучащей речи, автоматическое распознавание речи (ASR), Speech-to-Text (STT)-модель, State Of The Art (SOTA)-модель, облачное решение (Cloud-based), локальное решение (On-Premise), Docker-контейнер, русскоязычная речь

Введение. На сегодняшний день технологии преобразования речи в текст (STT) находят широкое применение в различных цифровых продуктах и сервисах: в социальных сетях для автоматической расшифровки голосовых сообщений [1], в голосовых ассистентах, опирающихся на STT для интерпретации запросов в режиме реального времени и обеспечения взаимодействия «человек–машина» на естественном языке [2]. С учетом активного роста исследований в области ASR и по данным отраслевого анализа, объем рынка голосовых технологий превысил 20 млрд USD в 2023 г. и продолжает расти с годовым темпом около 14,6 % за счет высокой потребности проектов в функционале автоматической транскрипции [3].

В рамках проекта «Виртуальный помощник студента» (чат-бот «Вопрошалыч»), развернутого в информационном пространстве Тюменского государственного университета (ТюмГУ), с целью повышения доступности сервиса для студентов, улучшить пользовательский опыт за счет естественного способа взаимодействия и снизить когнитивную нагрузку, связанную с вводом текста, возникла необходимость в добавлении возможности обработки голосовых сообщений от студентов.

Проведенный анализ выявил доступные подходы к внедрению в существующий проект возможности голосовой обработки: облачные сервисы (Cloud-based) и локальные (On-Premise) решения [4].

Облачные STT-сервисы обрабатывают аудиоданные на удаленных серверах провайдеров, что обеспечивает высокую масштабируемость и доступ к современным моделям, однако увеличивает риски утечек и нарушений нормативов. К ключевым системам можно отнести Google Cloud Speech-to-Text и Microsoft Azure Speech, а также сервисы AWS (Amazon Transcribe), AssemblyAI и др. Обычно они предоставляют готовые API: достаточно отправить аудиопоток или файл и получить результат транскрипции.

Локальные (On-Premise) системы распознавания речи, основанные на открытых фреймворках, выполняют обработку внутри инфраструктуры организации. Например, Kaldi — широко известный открытый инструмент для ASR, Vosk — легковесный офлайн-распознаватель речи с поддержкой более 20 языков. Аналогично, проект Mozilla DeepSpeech предлагает open-source сервис, способный работать на устройствах от Raspberry Pi до GPU-серверов. Подобные системы обеспечивают конфиденциальность, автономность, независимость от интернет-соединения и низкую задержку, однако не демонстрируют сопоставимого с современными

облачными моделями уровня точности распознавания. Kaldi демонстрирует «патологически плохой» WER во многих доменах [5], а Vosk[1] показывает ошибку свыше 16 % при транскрипции аудиокниг и до 40 % для телефонных разговоров. В то же время современные SOTA-модели, например, Whisper-Large-v3[2], дообученные на русском корпусе, достигают WER ниже 6,4 %, однако их эксплуатация предполагает завершение следующих процессов:

- развертывание и интеграция моделей распознавания речи;
- масштабирование вычислительных ресурсов для обработки растущей нагрузки;
- обеспечение высокой скорости работы (часто с использованием GPU);
- построение очередей обработки аудиопотоков или аудиофайлов;
- поддержка потокового распознавания (streaming ASR) для длительных аудио;
- организация интерфейсов (REST/API) и/или пользовательских приложений (GUI).

Проблема исследования. Готовые к эксплуатации локальные STT-системы, такие как Kaldi и Vosk, уступают современным SOTA-моделям в распознавании русскоязычной речи. Однако развертывание таких моделей в локальной среде требует обеспечения разработчиком мощной вычислительной инфраструктуры, что существенно ограничивает их практическое применение в условиях ограниченного бюджета и технической экспертизы.

Целью работы является создание контейнеризированного On-Premise решения для преобразования русскоязычной речи в текст с REST-интерфейсом, сочетающее легкость развертывания, использование Open-Source компонентов для минимизации затрат, локальную обработку аудио без выхода в публичные сети и выбор исходной модели по результатам сравнительного анализа WER/CER на русскоязычных корпусах для обеспечения высокого качества распознавания.

Материалы и методы. В качестве исходной модели обработки звучащей речи, была выбрана GigaAM-RNNT-v2[3] — «state of the art» (передовая) открытая акустическая модель от Sber, основанная на Conformer-энкодере, обученная на 50 тысячах часов разнообразных русскоязычных данных на основе подхода HuBERT и дообученная на задачу автоматического распознавания речи.

Основной метрикой качества систем автоматического распознавания речи (ASR) является Word Error Rate (WER), которая рассчитывается как отношение суммы замен, удалений и вставок слов к общему числу слов в эталонной транскрипции с помощью расстояния Левенштейна [6]. WER показывает долю ошибок в распознавании и используется для оценки и сравнения моделей.

Модель GigaAM-RNNT-v2 была выбрана на основании официальных результатов бенчмарка[4], подготовленного командой разработчиков, в котором сравниваются показатели данной модели с открытыми архитектурами Whisper-large-v3 и NeMo Conformer-RNNT по метрике Word Error Rate, рассчитанной на 8 крупных открытых параллельных корпусах рус-

[1] VOSK: сайт. — URL: https://alphacephei.com/vosk/

[2] antony66. Whisper Large V3 Russian // Hugging Face. — URL: https://huggingface.co/antony66/whisper-large-v3-russian

[3] GigaAM (Giga Acoustic Model) // GitHub. — URL: https://github.com/salute-developers/GigaAM

[4] GigaAM. Метрики качества: Word Error Rate // GitHub. — URL: https://github.com/salute-developers/GigaAM/blob/main/README_ru.md#метрики-качества-word-error-rate

скоязычных аудиоданных. На всех тестовых наборах модель GigaAM-RNNT-v2 демонстрирует наименьший показатель WER: для Golos Crowd — 2,2 %, для Golos Farfield — 3,9 %, для OpenSTT YouTube — 13,3 %, для OpenSTT Phone Calls — 20,0 %, для OpenSTT Audiobooks — 10,2 %, для Mozilla Common Voice 12 — 1,8 %, для Mozilla Common Voice 19 — 2,7 % и для Russian LibriSpeech — 5,5 %. Это свидетельствует о минимальной доле неправильно распознанных слов по сравнению с конкурирующими решениями.

Дополнительная метрика Character Error Rate (CER), измеряет точность распознавания на уровне отдельных символов, а не слов [7]. CER обеспечивает более детальный взгляд на точность распознавания, особенно в случаях, когда модель может правильно распознать большую часть слова, но ошибиться в нескольких символах.

С целью подтвердить оправданность выбора модели проведена оценка качества автоматической транскрипции голосовых сообщений. Для этого осуществлен сравнительный анализ аудиоданных, собранных при взаимодействии студентов с тестовой версией чат-бота «Вопрошалыч».

В тестировании приняли участие 25 студентов 2 курса бакалавриата ИТ-направлений ТюмГУ, которые направляли голосовые сообщения в адрес чат-бота. В результате была сформирована выборка, включающая 110 аудиофрагментов со средней продолжительностью 8 секунд: минимальная продолжительность — 2 секунды, максимальная — 15 секунд. Запросы студентов отражали типичные вопросы (об образовательных услугах, порядке подачи заявлений и др.), ответы на которые содержатся в корпоративном хранилище документов ТюмГУ. Для создания эталонной текстовой выборки, необходимой для оценки качества транскрибирования, участники предварительно вводили текстовую расшифровку своих голосовых сообщений.

Обработка и транскрипция аудиозаписей осуществлялись локально на вычислительной платформе (процессор AMD Ryzen 7 3700X, 16 ГБ оперативной памяти стандарта DDR4).

Для оценки качества распознавания речи использовалась библиотека jiwer, с помощью которой были рассчитаны ключевые метрики: Word Error Rate (WER) и Character Error Rate (CER) (табл. 1).

Таблица 1

Метрики качества (Word Error Rate, Character Error Rate, Speed)

Model	*Parameters*	*Word Error Rate (WER)*	*Character Error Rate (CER)*	*Speed (x = 1 second)*
Whisper-large-v3	1.5B	6.86	3.25	0.83x
vosk-model-small-ru-0.22	45M	5.67	1.71	0.231x
h2oai/faster-whisper-large-v3-turbo	809M	2.98	0.76	0.44x
GigaAM-CTC-v1	242M	6.53	2.63	0.566x
GigaAM-RNNT-v1	243M	3.03	0.75	0.476x
GigaAM-CTC-v2	242M	3.1	0.71	0.406x
GigaAM-RNNT-v2	243M	2.94	0.71	0.383x

Аудиоданные варьировались по качеству: в ряде записей присутствовали посторонние шумы (фоновая речь, уличный шум), что могло оказывать влияние на точность распознавания. В остальном записи отличались стандартной речевой интонацией и не содержали выраженных диалектных особенностей.

Полученные значения метрик WER и CER оказались минимальными по сравнению с аналогичными известными моделями и составили 2,94 % и 0,71 % соответственно, что свидетельствует о более высокой точности распознавания речи по сравнению с аналогами. Это подтверждает целесообразность выбора модели для реализации сервиса преобразования русскоязычной речи в текст.

Выбор архитектуры сервиса на основе REST API в контейнеризованной среде обусловлен разделением ответственности между клиентской и серверной частями, безсессионностью (statelessness) запросов и наличием унифицированного интерфейса, что упрощает взаимодействие компонентов и повышает масштабируемость системы [8].

Результаты. В рамках проекта «Виртуальный помощник» был разработан и интегрирован сервис автоматического преобразования речи в текст, который был протестирован на вычислительной платформе, оснащенной системой на кристалле Apple M4, включающей 10-ядерный центральный процессор, конфигурация устройства включала 24 ГБ унифицированной памяти LPDDR5X с пропускной способностью до 120 ГБ/с.

Сервис реализован в виде RESTful API, развернутого в Docker-контейнере, что обеспечивает стандартизированный, независимый от платформы способ взаимодействия с другими компонентами системы.

В качестве модели распознавания речи используется GigaAM-RNNT-v2, оптимизированная для инференса на центральных процессорах (CPU), что позволяет использовать доступные вычислительные ресурсы без необходимости в специализированном оборудовании.

Взаимодействие компонентов "Виртуального помощника", происходит следующим образом: пользователь отправляет голосовое сообщение, микросервис Chatbot сохраняет аудиофайл, конвертирует в унифицированный формат WAV 16 кГц и пересылает по REST-API в сервис распознавания речи (см. рис. 1).

Функциональность сервиса включает POST-эндпоинт, который принимает аудиофайл, передает его в акустический модуль и формирует JSON-ответ с транскрибированным текстом. Полученный JSON возвращается в микросервис Chatbot, где обрабатывается аналогично обычному текстовому сообщению и далее направляется в последующие компоненты системы.

На данный момент микросервис распознавания речи предназначен для обработки коротких аудиозаписей продолжительностью до 30 секунд.

Заключение. В рамках проекта «Виртуальный помощник» был разработан, интегрирован и протестирован микросервис Chatbot обработки звучащей русскоязычной речи, подтвердивший возможность On-Premise обработки аудиоданных с использованием открытых STT-моделей для автоматического распознавания речи на русском языке.

Проведенная разработка микросервиса Chatbot выявила ключевые направления для дальнейшего развития решения — внедрение возможности обработки аудиозаписей любой продолжительности, проведение нагрузочного тестирования, а также открытие исходного кода проекта (Open-Source).

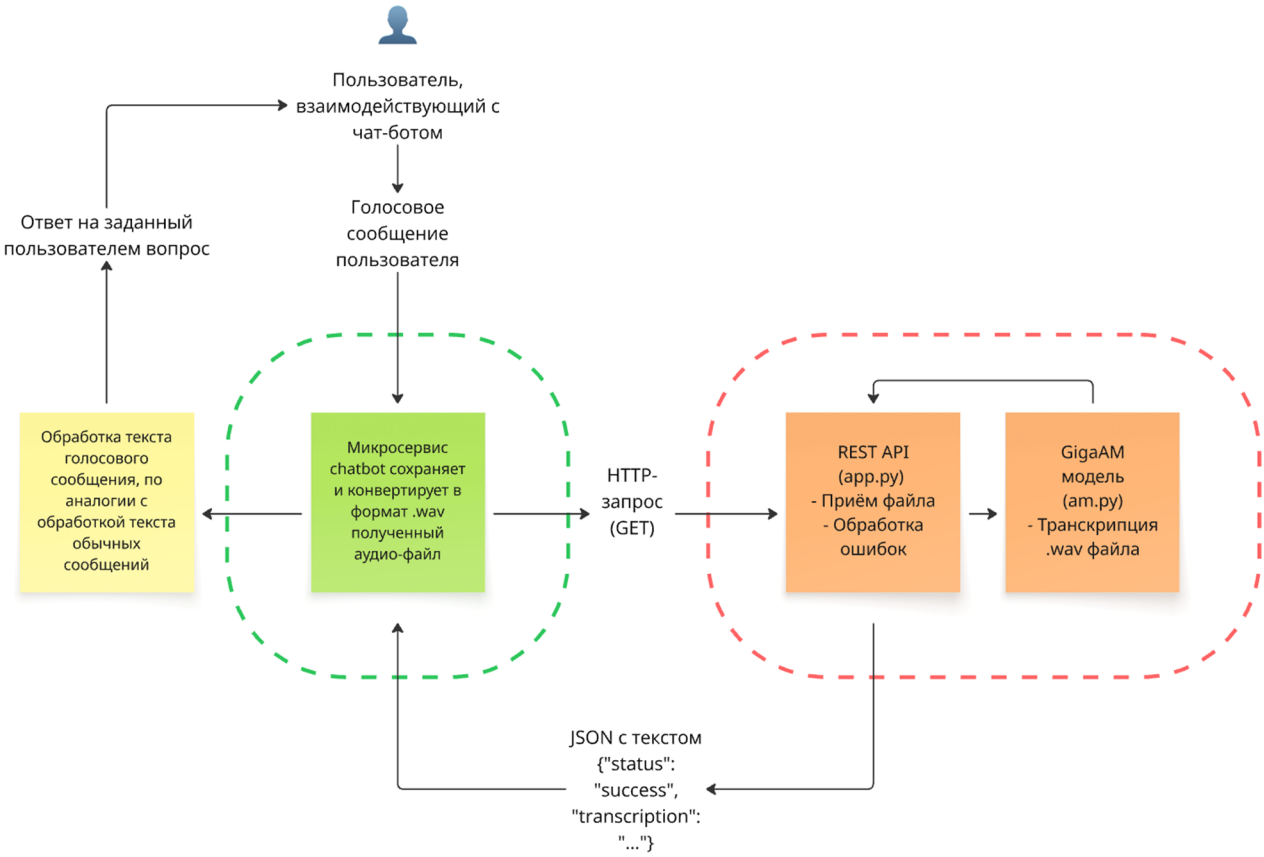

Рис. 1. Интеграция разработанного решения в микросервисную архитектуру проекта "Вопрошалыч"

СПИСОК ИСТОЧНИКОВ

1. Automatic speech-to-text transcription: evidence from a smartphone survey with voice answers // Taylor & Francis Online. URL: https://www.tandfonline.com/doi/epdf/10.1080/13645579.2024.2443633? needAccess=true (дата обращения: 04.05.2025).
2. Speech AI for All: Promoting Accessibility, Fairness, Inclusivity, and Equity // HESP, University of Maryland. URL: https://hesp.umd.edu/sites/hesp.umd.edu/files/pubs/CHI25_Speech_AI_workshop.pdf (дата обращения: 04.05.2025).
3. Voice And Speech Recognition Market Size, Share & Trends Analysis Report By Function (Speech Recognition, Voice Recognition), By Technology, By Vertical, By Region, And Segment Forecasts, 2024–2030 // Grand View Research. URL: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market (дата обращения: 04.05.2025).
4. Исследование возможности внедрения голосовых ассистентов на объекты добычи и переработки нефти // CyberLeninka. URL: https://cyberleninka.ru/article/n/issledovanie-vozmozhnosti-vnedreniya-golosovyh-assistentov-na-obekty-dobychi-i-pererabotki-nefti/viewer (дата обращения: 14.05.2025).
5. Studying the Practices of Deploying Machine Learning Projects on Docker // ACM Digital Library. URL: https://dl.acm.org/doi/10.1145/3530019.3530039 (дата обращения: 04.05.2025).
6. Гордеева Л., Ершов В., Лабутин И., Кураленок И. Meaning Error Rate // CEUR Workshop Proceedings. 2020. Vol-2691. С. 1–8.
7. Zhang Y, Wang X, He J, Liu Y. Advocating Character Error Rate for Multilingual ASR Evaluation // arXiv preprint arXiv:2401.11700. 2024. URL: https://arxiv.org/pdf/2401.11700 (дата обращения: 14.05.2025).

8. Fielding RT. Architectural Styles and the Design of Network-based Software Architectures: Doctoral dissertation, University of California, Irvine. 2000. URL: https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm (дата обращения: 04.05.2025).
9. Benchmarking Top Open Source Speech Recognition Models: Whisper, Facebook wav2vec2, and Kaldi // Deepgram. URL: https://deepgram.com/learn/benchmarking-top-open-source-speech-models (дата обращения: 04.05.2025).
10. Pahl C, Lee B. Containers and Cloud: From LXC to Docker to Kubernetes // Proceedings of the 2015 IEEE International Conference on Cloud Computing Technology and Science (CloudCom). 2015. P. 799–805. DOI: 10.1109/CloudCom.2015.174 (дата обращения: 04.05.2025).
11. Transforming industrial automation: voice recognition control via containerized PLC device // Scientific Reports. URL: https://www.nature.com/articles/s41598-024-81172-w (дата обращения: 04.05.2025).