СТАТЬЯ

УДК 004.934

ОЦЕНКА КАЧЕСТВА РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ НА ЧИСТЫХ И ЗАШУМЛЕННЫХ АУДИОДАННЫХ

Тукаев В.Р., Беляева М.Б.

ФГБОУ ВО «Уфимский университет науки и технологий», Стерлитамакский филиал, Стерлитамак, e-mail: vildantukaev@yandex.ru

Проблема автоматического распознавания русской речи в условиях шума остается актуальной задачей для многих прикладных систем, требующих точной обработки аудиоданных в реальных условиях. Целью исследования стала оценка и сравнительный анализ качества распознавания русской речи четырьмя современными моделями: Whisper-large-v3, Vosk-model-ru-0.42 и двумя вариантами модели GigaAM второй версии с различными функциями потерь: на основе временной классификации с соединениями и на основе рекуррентной нейронной сети преобразователя. Исследование проводилось на чистых и искусственно зашумленных аудиоданных, подготовленных на основе открытого набора данных с добавлением шумов из датасета аудиозаписей окружающей среды при определенном соотношении сигнал/шум. Качество распознавания оценивалось с использованием метрик: коэффициентов ошибок на уровне слов и символов, метрики ошибок сопоставления и метрики потери информации на уровне слов. Экспериментальная часть включала обработку выборки из 1000 аудиозаписей различной длительности, содержащих как спонтанную речь, так и подготовленные тексты. Результаты исследования показали, что модели GigaAM обеспечили наилучшее качество распознавания речи как в чистых, так и в зашумленных условиях. Особенно выделилась модель GigaAM с функцией потерь на основе рекуррентной нейронной сети преобразователя, которая продемонстрировала оптимальное сочетание точности и устойчивости к шуму. Полученные результаты имеют практическую ценность для выбора оптимальной модели распознавания речи в различных прикладных задачах, связанных с обработкой русской речи в реальных условиях.

Ключевые слова: автоматическое распознавание речи, точность распознавания речи, показатели оценивания, акустический шум, русская речь, сравнительный анализ

EVALUATION OF RUSSIAN SPEECH RECOGNITION QUALITY ON CLEAN AND NOISY AUDIO DATA

Tukaev V.R., Belyaeva M.B.

[1]Sterlitamak branch of the Ufa University of Science and Technology, Sterlitamak, e-mail: vildantukaev@yandex.ru

The problem of automatic recognition of Russian speech in noisy environments remains a pressing issue for many applied systems that require accurate audio data processing in real-world conditions. The aim of the study was to evaluate and compare the quality of Russian speech recognition by four modern models: Whisper-large-v3, Vosk-model-ru-0.42, and two variants of the second version of the GigaAM model with different loss functions: based on temporal classification with connections and based on a recurrent neural network converter. The study was conducted on clean and artificially noisy audio data prepared on the basis of an open dataset with the addition of noise from a dataset of environmental audio recordings at a specific signal-to-noise ratio. Recognition quality was evaluated using metrics: word and character error rates, matching error metrics, and word-level information loss metrics. The experimental part included processing a sample of 1,000 audio recordings of varying lengths, containing both spontaneous speech and prepared texts. The results of the study showed that GigaAM models provided the best speech recognition quality in both clean and noisy conditions. The GigaAM model with a loss function based on a recurrent neural network converter stood out in particular, demonstrating an optimal combination of accuracy and noise resistance. The results obtained are of practical value for selecting the optimal speech recognition model in various application tasks related to the processing of Russian speech in real-world conditions.

Keywords: automatic speech recognition, speech recognition accuracy, evaluation indicators, acoustic noise, Russian speech, comparative analysis

Введение

Автоматическое распознавание речи (ASR) стало неотъемлемой частью современных технологий, от голосовых помощников и систем диктовки до автоматической расшифровки звукозаписей и обработки телефонных разговоров в службах поддержки [1, с. 5]. Точность и надежность моделей распознавания речи напрямую влияют на пользовательский опыт и определяют возможности их практического применения. Особую актуальность приобретает задача разработки и оценки таких моделей для различных языков.

Одной из ключевых проблем при использовании систем распознавания речи в реальных условиях является наличие фонового шума. Шумы различного происхождения (окружающая среда, технические помехи, речь других людей) могут значительно снижать точность преобразования речи из аудио в текст.

Целью данной работы является оценка и сравнительный анализ качества распознавания русской речи четырьмя современными ASR-моделями (Whisper-large-v3, Vosk-model-ru-0.42, GigaAM-CTC-v2, GigaAM-RNNT-v2) на чистых и зашумлённых аудиоданных.

Материалы и методы исследования

Исследование включает в себя подготовку тестового русскоязычного набора данных, выбор и описание современных ASR-моделей, а также определение метрик для оценки качества распознавания речи. Проведены эксперименты по преобразованию речи в текст на подготовленных данных с использованием выбранных моделей. Полученные результаты проанализированы для оценки и сравнения качества распознавания и устойчивости моделей к шуму. На основе результатов сформулированы выводы о применимости исследуемых моделей для распознавания русской речи в различных акустических условиях.

В качестве основного источника данных использовался открытый русскоязычный набор данных Silero Open STT (Sh1man/silero_open_stt) [2]. Для проведения экспериментов было отобрано подмножество buriy_audio_books_2, из которого был сформирован срез из 1000 аудиофайлов. Длительность выбранных аудиофайлов варьировалась от 0.34 до 17.97 секунд, со средней длительностью 2.24 секунды.

Для имитации реальных условий применения ASR-модели была создана зашумлённая версия тестового набора данных. Процесс добавления шума включал несколько последовательных этапов обработки аудиоматериала.

Первоначально каждая чистая аудиозапись из выборки приводилась к единому формату с частотой дискретизации 16 кГц и преобразовывалась в монофонический сигнал для стандартизации входных данных. В качестве источников акустических помех использовались аудиозаписи из набора данных ESC-50 [3], содержащего 2000 коротких записей различных звуков окружающей среды, распределенных по 50 классам (по 40 примеров в каждом).

Для каждой чистой записи алгоритмически выбирался случайный фрагмент шума из ESC-50, который затем добавлялся к исходному сигналу с фиксированным отношением сигнал/шум (SNR) на уровне 5 дБ. Данный уровень SNR был выбран как репрезентативный для условий умеренного акустического зашумления.

На заключительном этапе обработки производилась нормализация амплитуды полученного зашумлённого сигнала для предотвращения искажений. В результате для каждой из 1000 исходных аудиозаписей были подготовлены две версии: оригинальная (чистая) и с добавленным шумом.

В исследовании сравнивались четыре модели ASR.

Whisper-large-v3 – многоязычная модель от OpenAI, основанная на архитектуре Transformer [4]. Данная модель обучена на обширном наборе разнообразных аудиоданных (680 тыс. часов), включая значительную долю неанглийской речи. Версия large-v3 является наиболее производительной в семействе Whisper.

Vosk-model-ru-0.42 – распространенная ASR-модель с открытым исходным кодом, основанная на технологиях Kaldi [5]. Модель адаптирована для русского языка и оптимизирована для функционирования на различных вычислительных платформах.

GigaAM-CTC-v2 – модель из семейства Giga Acoustic Model (GigaAM), разработана для русского языка. Построена на архитектуре Conformer и предобучена с использованием методов самоконтролируемого обучения (в частности, HuBERT) на корпусе, превышающем 50 000 часов русской речи. Данная версия дообучена с применением функции потерь Connectionist Temporal Classification (CTC) [6].

GigaAM-RNNT-v2 – вторая модель из семейства GigaAM, также основана на архитектуре Conformer и предобучена с использованием алгоритма HuBERT [7]. Ключевое отличие состоит в применении функции потерь RNN Transducer (RNNT) [8] для дообучения.

Модели GigaAM позиционируются как решения с повышенной устойчивостью к акустическим помехам [9].

Для количественной оценки качества распознавания речи использовались четыре стандартные метрики [10]:

WER (Word Error Rate) – коэффициент ошибок на уровне слов. Наиболее распространенная метрика, измеряющая процент неправильно распознанных слов.

$$WER = \frac{(S + D + I)}{T} \times 100\%,$$

где S – число замен, D – число удалений, I – число вставок, T – число слов в эталонной транскрипции.

CER (Character Error Rate) – коэффициент ошибок на уровне символов. Аналогичен WER, но вычисляется на уровне символов. Метрика полезна для языков с богатой морфологией и для оценки точности

распознавания имен собственных или редких слов.

$$CER = \frac{\left(S_{char} + D_{char} + I_{char}\right)}{T_{char}} \times 100\%,$$

где T_{char} – число символов эталонной транскрипции.

MER (Match Error Rate) – метрика, учитывающая не только ошибки (замены, удаления, вставки), но и правильно распознанные слова (совпадения). Рассчитывается как отношение суммы ошибок к сумме совпадений и замен. Иногда интерпретируется как доля «неправильной» информации в распознанном тексте относительно «правильной».

$$MER = \frac{\left(S + D + I\right)}{\left(H + S + D + I\right)} \times 100\%,$$

где H – число совпадений.

WIL (Word Information Lost) – метрика, которая измеряет потерю информации при распознавании речи. В отличие от WER, которая просто подсчитывает процент ошибок, WIL пытается оценить, насколько важная информация была потеряна или искажена.

$$WIL = \left(1 - \frac{H^2}{\left(T * T_0\right)}\right) \times 100\%,$$

где T – число слов в эталонной транскрипции, T_0 – число слов в гипотезе распознавания (в результате работы ASR-модели).

Для всех метрик (WER, CER, MER, WIL) меньшее значение указывает на лучшее качество преобразования речи из аудио в текст.

Все эксперименты проводились в облачной среде Google Colaboratory. Измерение времени выполнения распознавания для сравниваемых систем не проводилось из-за существенных различий в возможностях аппаратного ускорения. Модели Whisper и GigaAM поддерживают GPU-ускорение через CUDA, тогда как Vosk-model-ru-0.42 такой поддержкой не обладает и выполнялась исключительно на центральном процессоре (CPU). Прямое сопоставление временных показателей моделей привело бы к некорректным выводам. Основной акцент исследования был сделан на анализе точности распознавания речи.

Результаты исследования и их обсуждение

В ходе исследования были получены значения метрик WER, CER, MER и WIL для каждой из четырёх моделей на чистом (Ч) и зашумлённом (Ш) наборах данных. Результаты сведены в таблицу и представлены на рисунках 1-4.

Модель Whisper-large-v3 показала наихудшие результаты по метрике WER: 65.3% на чистых данных и 67.4% на зашумлённых (разница 2.1%).

У модели Vosk-model-ru-0.42 наблюдалось значительное снижение качества: WER вырос с 16.4% до 28.2%, что составляет разницу в 11.8%.

Модель GigaAM-CTC-v2 продемонстрировала WER 16.6% на чистых и 22.7% на зашумлённых данных (разница 6.0%).

Наилучшие показатели были у GigaAM-RNNT-v2: 15.5% на чистых и 21.8% на зашумлённых данных (разница 6.3%).

Таким образом, GigaAM-RNNT-v2 является лидером по метрике WER в обоих условиях (рис. 1).

Результаты по метрике CER согласуются с предыдущими наблюдениями.

Whisper-large-v3 снова показала самые высокие значения ошибки: 23.3% на чистых и 24.9% на зашумлённых данных (разница 1.6%).

У модели Vosk-model-ru-0.42 CER увеличился с 4.6% до 12.9%, показав существенное падение качества (разница 8.3%).

Результаты оценки качества распознавания речи моделей ASR (значения в %)

Модель	Данные	WER	CER	MER	WIL
Whisper-large-v3	Ч	65.3	23.3	58.6	74.0
	Ш	67.4	24.9	62.7	77.1
Vosk-model-ru-0.42	Ч	16.4	4.6	15.3	20.8
	Ш	28.2	12.9	27.1	34.6
GigaAM-CTC-v2	Ч	16.6	4.4	15.6	21.1
	Ш	22.7	8.4	21.8	28.0
GigaAM-RNNT-v2	Ч	15.5	4.5	14.5	19.5
	Ш	21.8	8.6	20.7	26.4

Примечание: таблица составлена авторами на основе полученных данных в ходе исследования.

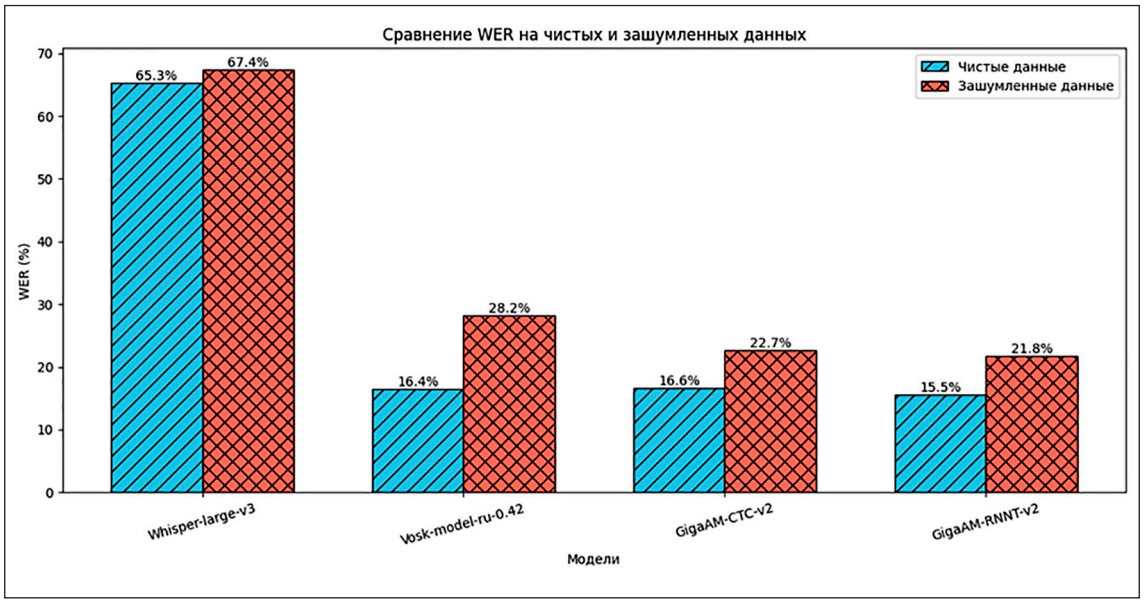

Рис. 1. Сравнение WER на чистых и зашумленных данных
Источник: составлен авторами по результатам данного исследования

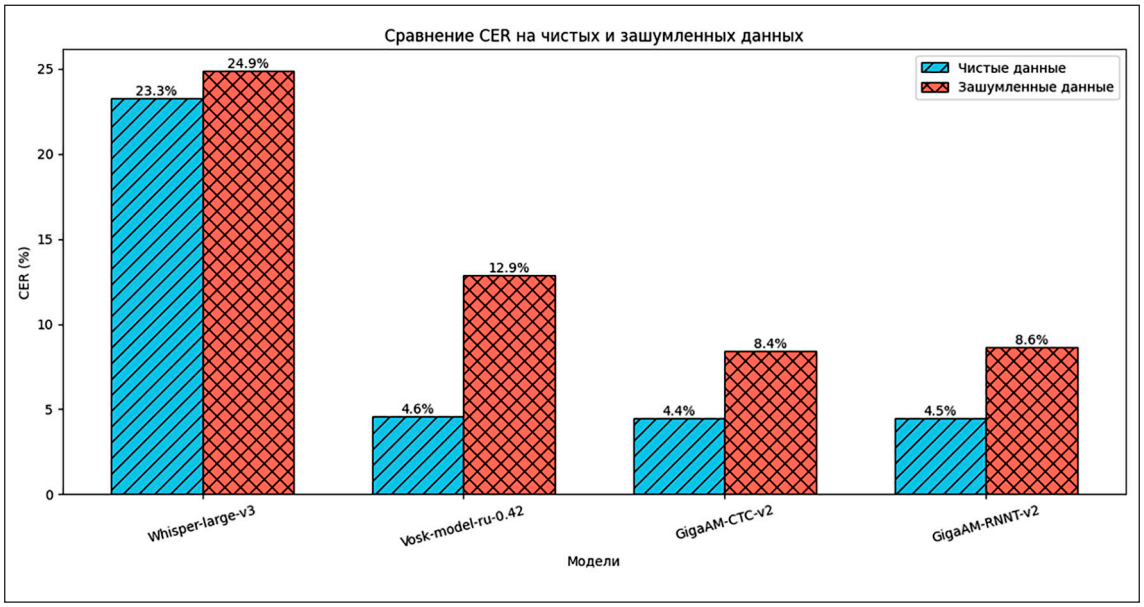

Рис. 2. Сравнение CER на чистых и зашумленных данных
Источник: составлен авторами по результатам данного исследования

Модель GigaAM-CTC-v2 продемонстрировала наилучшие результаты по этой метрике: 4.4% на чистых и 8.4% на зашумлённых данных (разница 4.0%).

GigaAM-RNNT-v2 показала близкие значения: 4.5% и 8.6% соответственно (разница 4.1%).

По метрике CER лидирует GigaAM-CTC-v2, но разница с GigaAM-RNNT-v2 незначительна (рис. 2).

Метрика MER подтверждает общую картину качества распознавания моделей (рис. 3).

Whisper-large-v3 снова показала наивысший уровень ошибок: 58.6% на чистых и 62.7% на зашумлённых данных (разница 4.1%).

Модель Vosk-model-ru-0.42 также продемонстрировала значительное ухудшение: MER вырос с 15.3% до 27.1% (разница 11.8%).

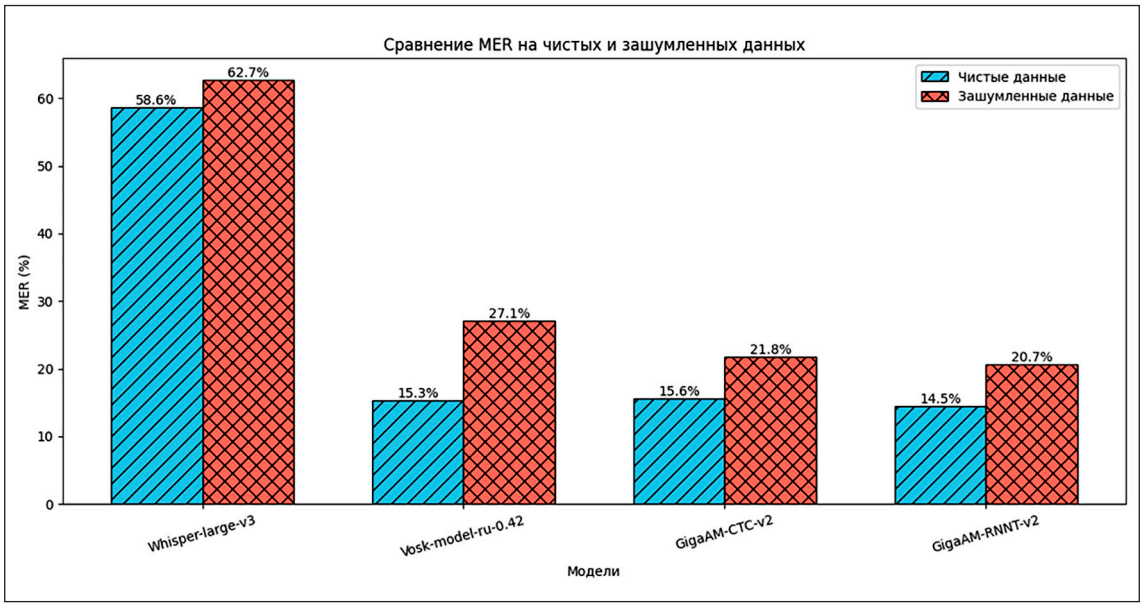

Рис. 3. Сравнение MER на чистых и зашумленных данных
Источник: составлен авторами по результатам данного исследования

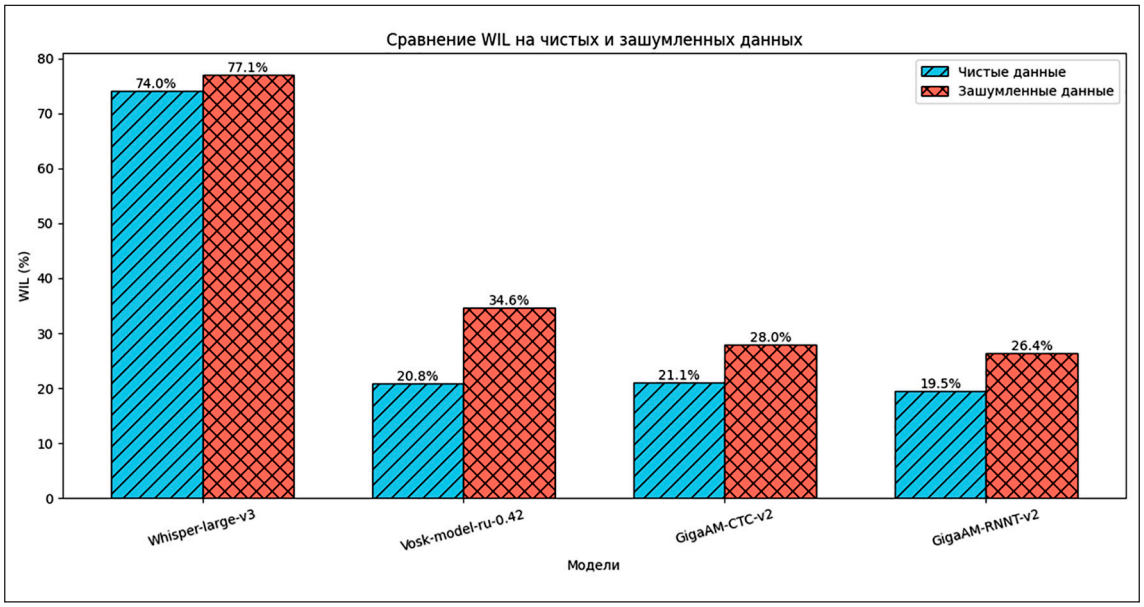

Рис. 4. Сравнение WIL на чистых и зашумленных данных
Источник: составлен авторами по результатам данного исследования

Модели GigaAM показали лучшие и более стабильные результаты. GigaAM-CTC-v2: 15.6% на чистых и 21.8% на зашумлённых (разница 6.2%). GigaAM-RNNT-v2 снова оказалась лучшей: 14.5% на чистых и 20.7% на зашумлённых данных (разница 6.3%).

При оценке по метрике WIL выявлено преимущество моделей GigaAM (рис. 4).

Whisper-large-v3 показала высокие значения WIL: 74.0% на чистых и 77.1% на зашумлённых данных (разница 3.1%).

Vosk-model-ru-0.42 снова продемонстрировала наибольшую чувствительность к шуму: WIL вырос с 20.8% до 34.6% (разница 13.8%).

Модели GigaAM показали схожую устойчивость к шуму. Gi-gaAM-CTC-v2: 21.1% на чистых и 28.0% на зашумлённых (разница 6.9%). GigaAM-RNNT-v2 вновь показала лучшие значения: 19.5% на чистых и 26.4% на зашумлённых данных (разница 6.9%).

Заключение

Проведенное исследование позволяет сделать следующие выводы относительно качества распознавания современных моделей автоматического распознавания русской речи в различных акустических условиях.

Модели семейства GigaAM (GigaAM-CTC-v2 и GigaAM-RNNT-v2) продемонстрировали наиболее высокое качество распознавания как на чистых, так и на зашумлённых данных по всем рассматривае-мым метрикам. При этом GigaAM-RNNT-v2 показала небольшое, но стабильное преимущество над версией с CTC-потерями по ключевым метрикам WER, MER и WIL. Высокая эффективность этих моделей, вероятно, объясняется комбинацией архитектуры Conformer, большого объема специфичных для русского языка данных для предобучения (более 50 тыс. часов русской речи) и использованием методов самоконтролируемого обучения.

Модель Vosk-model-ru-0.42 продемонстрировала приемлемое качество на чистых данных, сопоставимое с моделями GigaAM по метрикам WER и MER, и даже превосходящее их по CER. Однако её качество существенно ухудшилось при добавлении шума. Разница в показателях WER, MER и WIL между чистыми и зашумлёнными данными для Vosk оказалась почти в два раза выше, чем для моделей GigaAM, что свидетельствует о меньшей устойчивости данной модели к акустическим помехам.

Модель Whisper-large-v3, несмотря на значительный размер и многоязычность, показала значительно более низкие результаты по всем метрикам на исследуемом русскоязычном наборе данных по сравнению с моделями, специально обученными или дообученными на обширном корпусе русской речи (Vosk, GigaAM). Возможными причинами такого результата могут быть недостаточная адаптация к особенностям русской речи, специфика обучающей выборки, либо особенности алгоритма декодирования. При этом данная модель показала относительно низкую чувствительность к шуму (разница в WER всего 2.1%).

Таким образом, результаты оценки и сравнительного анализа четырех современных ASR-моделей для русского языка (Whisper-large-v3, Vosk-model-ru-0.42, GigaAM-CTC-v2, GigaAM-RNNT-v2) на чистых и зашумлённых аудиоданных позволяют заключить, что модель GigaAM-RNNT-v2 демонстрирует оптимальное сочетание высокой точности распознавания и устойчивости к акустическим помехам. Полученные результаты имеют практическую ценность для выбора ASR-решений в различных прикладных задачах, связанных с обработкой русской речи в реальных условиях.

Список литературы

1. Тампель ИБ., Карпов А.А. Автоматическое распознавание речи: учебное пособие. СПб.: Университет ИТМО, 2016. 138 с. URL: https://books.ifmo.ru/file/pdf/2232.pdf (дата обращения: 26.05.2025).

2. Sh1man. Silero Open STT // Hugging Face. 2025. URL: https://huggingface.co/datasets/Sh1man/silero_open_stt (дата обращения: 26.05.2025).

3. Piczak, K.J. ESC-50: Dataset for Environmental Sound Classifi-cation // GitHub. URL: https://github.com/karolpiczak/ESC-50 (дата обращения: 26.05.2025).

4. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision // Proceedings of the 40th International Conference on Machine Learning. 2023. Vol. 202. P. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html (дата обращения: 26.05.2025).

5. Alpha Cephei. Vosk Speech Recognition Toolkit // Официальный сайт Vosk. URL: https://alphacephei.com/vosk/ (дата обращения: 26.05.2025).

6. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recur-rent neural networks // Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA. 2006. P. 369–376. DOI: 10.1145/1143844.1143891.

7. Hsu W.-N., Bolte B., Tsai Y.-H.H., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. Vol. 29. P. 3451–3460. DOI: 10.1109/TASLP.2021.3122291.

8. Graves A. Sequence Transduction with Recurrent Neural Networks // Proceedings of the ICML 2012 Workshop on Representation Learning. 2012. DOI: 10.48550/arXiv.1211.3711.

9. GigaAM: the family of open-source acoustic models for speech processing // GitHub. URL: https://github.com/salute-developers/GigaAM (дата обращения: 26.05.2025).

10. Карпов А.А., Кипяткова И.С. Методология оценивания работы систем автоматического распознавания речи // Известия высших учебных заведений. Приборостроение. 2012. № 55 (4). С. 15-21. URL: https://elibrary.ru/item.asp?id=18045933 (дата обращения: 26.05.2025). EDN: PEXJRL.