DOI: 10.15514/ISPRAS-2026-38(3)-49

Исследование влияния сетевых деградаций на модели распознавания речи

[1] *А.В. Полевой, ORCID: 0009-0000-2216-0468 <polevoianton@bk.ru>*
[1,2] *Н.В. Лукашевич, ORCID: 0000-0002-1883-4121 <louk_nat@mail.ru>*

[1] *Московский государственный университет им. М.В. Ломоносова, факультет ВМК, Россия, 119234, г. Москва, Ленинские горы, 1, стр. 4.*

[2] *Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова, Россия, 119234, г. Москва, Ленинские горы, 1, стр. 4.*

Аннотация. Несмотря на успехи моделей автоматического распознавания речи на различных наборах данных и языках, применение моделей в повседневной жизни не позволяет использовать их в различных сценариях, например, звонки с нестабильным сетевым соединением или телефонные каналы с помехами. В данной работе был разработан и представлен специализированный тестовый набор русскоязычной речи, ключевой особенностью которого является репрезентативный набор данных с контролируемыми деградациями сигнала, вызванными нестабильным интернет-соединением. На предложенном наборе была проведена апробация и сравнительный анализ современных подходов к распознаванию речи. Для количественной оценки степени искажений использовался автоматизированный метод, основанный на анализе совокупности акустических характеристик сигнала и нейросетевых метрик. Полученные результаты позволяют выявить методы, наиболее устойчивые к акустическим деградациям.

Ключевые слова: автоматическое распознавание речи; автоматическое прогнозирование WER; аудио тестовый набор речевых записей; аудиозаписи с нестабильным интернет-соединением.

Для цитирования: Полевой А.В., Лукашевич Н.В. Исследование влияния сетевых деградаций на модели распознавания речи. Труды ИСП РАН, том 38, вып. 3, часть 4, 2026 г., стр. 101–118. DOI: 10.15514/ISPRAS–2026–38(3)–49.

Research on the impact of network degradation on automatic speech recognition models

[1] *A.V. Polevoi, ORCID: 0009-0000-2216-0468 <polevoianton@bk.ru>*
[1,2] *N.V. Loukachevitch, ORCID: 0000-0002-1883-4121 <louk_nat@mail.ru>*

[1] *Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, building 4, 1, Leninskie Gory, 119234, Moscow, Russia.*

[2] *Lomonosov Moscow State University, Research Computing Center, building 4, 1, Leninskie Gory, 119234, Moscow, Russia.*

Abstract. Despite the success of automatic speech recognition models on various datasets and in different languages, their use in everyday life is still limited due to their inability to handle certain scenarios, such as calls with unstable network connections or telephone channels with interference. In this paper, we present a specialized benchmark for Russian-language speech, which includes a representative dataset that simulates the effects of an unstable internet connection on speech. This benchmark was designed to test and compare the performance of modern speech recognition approaches. To quantify the level of degradation, we used an automated method based on analyzing a set of acoustic features and neural network metrics. The results obtained from our benchmark allow us to identify methods that are more resistant to acoustic distortion. These methods can be used to improve the reliability of speech recognition systems in real-world scenarios with challenging conditions.

Keywords: automatic speech recognition; WER estimation; audio benchmark of speech recordings; audio recordings with unstable internet connection.

For citation: Polevoi A.V., Loukachevitch N.V. Research on the impact of network degradation on automatic speech recognition models. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 3, part 4, 2026, pp. 101-118 (in Russian). DOI: 10.15514/ISPRAS-2026-38(3)-49.

1. Введение

Развитие систем автоматического распознавания речи позволяет использовать модели для различных сценариев использования: автоматическая работа колл-центров [1], голосовое управление на предприятиях [2], в домах и автомобилях [3], протоколирование конференций и совещаний [4] и другие. Тем не менее, несмотря на стабильные результаты для сигналов известного качества (ручные микрофоны, запись в помещениях со стабильным интернет-соединением), модели показывают нестабильные результаты в присутствие акустических деградаций [5-6].

Под акустической деградацией будем понимать искажение исходного речевого сигнала, без изменения первоначального текста. Традиционно, выделяют несколько типов акустических искажений [7]:

- Реальные фоновые шумы различной природы: шум мотора, различных предприятий и заводов, городского транспорта и аналогичные;
- Синтетические (сгенерированные) шумы: белый, розовый и аналогичные;
- Реверберация внутри зданий и помещений (также возможна реверберация вне помещений, например, в горной местности, но встречается реже): высокие потолки, гулкие стены и аналогичные;
- Дефекты устройств звукозаписи: амплитудное ограничение сигнала (clipping), отсутствие нормализации динамического диапазона;
- Сжатие сигналов при передаче и потери, вносимые нестабильным интернет-соединением.

При этом деградации на практике могут носить как временный, так и постоянный характер действия. Кроме того, несмотря на обилие наборов данных для систем распознавания речи, отсутствуют контролируемые тестовые наборы, построение которых позволит выбирать модели опираясь на конкретные требования.

Наша работа направлена на систематическое исследование проблем, связанных с распознаванием речевых сигналов с акустическими деградациями, вызванными нестабильным интернет-соединением:

- Построение актуального тестового набора данных русской речи с контролируемым характером акустических деградаций из-за нестабильного интернет-соединения;
- Апробация современных подходов автоматического распознавания речи на предложенном наборе;
- Автоматическая оценка деградаций при помощи подхода на основе совокупности акустических характеристик сигналов и автоматических нейросетевых метрик;
- Поиск перспективных комбинированных с большими языковыми моделями методов для улучшения качества распознавания в предложенном сценарии.

2. Обзор литературы

В этом разделе начнем с описания различных подходов для автоматического распознавания речи и трудностях в распознавании, которые данные подходы имеют. Существующие исследования направлены на исследование адаптации существующих архитектур моделей распознавания речи под выбранный домен.

2.1 Современные подходы распознавания речи

Наиболее популярной архитектурой для распознавания является Conformer [8]. Данная архитектура предоставляет возможность гибкого обучения и настройки под конкретную задачу. К основным плюсам данной архитектуры стоит отнести небольшое количество параметров (compute-optimal), трансформерная модель, адаптированная под работу с речевыми сигналами, а также скорость вычислений и возможность применения архитектуры Conformer к так называемой вариации Streaming Cache Aware, когда модель работает со звуковым потоком напрямую, поддерживая при этом внутреннее состояние. Существуют многочисленные улучшения и оптимизации Conformer. В частности:

- Zipformer – ускоренный и более эффективный энкодер [9]. Авторы вводят новые функции нормализации (BiasNorm) и активации (Swoosh), что позволяет достичь сопоставимых или лучших результатов, чем у исходной модели Conformer, при двухкратном ускорении обучения;
- Squeezeformer [10] – модификация исходной архитектуры Conformer: центральные слои работают на пониженной частоте дискретизации, а блок обработки упрощается до вида исходной модели. Это позволяет ускорить расчет при минимальной потере качества;
- Branchformer [11] – расширяет Conformer за счет параллельных ветвей самовнимания, что даёт модели возможность анализировать аудиосигнал разными способами одновременно;
- FastConformer [12] – вариант с уменьшенными сверточными ядрами и дополнительной оптимизацией. Он примерно в 2.4 раза быстрее классического Conformer при минимальном ухудшении качества распознавания.

Существует также и подходы для получения единой (Foundational) универсальной модели, например модели семейства Whisper (OpenAI [13]). Авторы обучают модель сразу нескольким задачам, такие как: определение конкретного языка анализируемого фрагмента, выделение временных меток, а также возможности перевода на английский язык (если декодеру подается метка «translate»). При этом они используют, в отличие от архитектуры Conformer частично или даже неполно размеченные данные (weakly-supervised), взятые из субтитров. Поэтому зачастую модель страдает различного рода галлюцинациями на тишину или какие-то незначительные шорохи, например, «Спасибо» [14] для русского языка.

В настоящее время наиболее активно развиваются подходы, целиком использующие большие языковые модели (Large Language Models, LLM), так называемые речевые большие языковые модели, SpeechLLM. Основная идея данной группы методов заключается в адаптации выходного пространства аудио энкодера для текстовой модели. Идея в том, чтобы адаптировать выход аудио-энкодера к входному пространству текстовой LLM, позволяя модели «понимать» речь как ещё один модуль. Примеры таких систем:

- LLaMA-Omni [15] – объединяет предобученный аудио-энкодер, адаптер, языковую модель и потоковый декодер, благодаря чему может одновременно генерировать текст и голосовой ответ по устному запросу, не требуя явной промежуточной транскрипции;
- AudioPaLM [16] – объединяет текстовую модель PaLM-2 и аудио модель AudioLM в единую мультимодальную архитектуру, достигающую высоких результатов в задачах распознавания и перевода речи;
- Qwen3-Omni [17] – мультимодальная модель, демонстрирующая лучшие на различных наборах данных. Она обрабатывает как текст, так и аудио: поддерживает распознавание речи длиной до 40 минут, понимает свыше 100 языков и в целом превосходит коммерческие аналоги (например, Gemini-2.5-Pro, GPT-4o-Transcribe) по точности распознавания.

Авторы таких моделей на основе LLM обещают унифицировать обработку текста и звука, однако их практическое применение и стабильность результатов активно исследуется.

2.2 Акустические деградации: обзор

В реальных условиях речь часто искажается фоновым шумом и реверберацией. Естественные шумы (фоновые звуки улицы, офиса, толпы) и искусственные шумы (синтетический шум, наложенный на сигнал) снижают разборчивость. Реверберация – многократное отражение звука в помещении – также искажает спектр речи: для разных помещений (размер, длинный коридор) качество распознавания речи может изменяться в широких пределах. Например, в корпусе CHiME-5 записаны диалоги в 20 реальных домах с бытовыми шумами (кухня, кондиционер и так далее) и разной акустикой [18]. Аналогично в VOiCES данные записаны в двух мебелированных комнатах с фоновыми шумами и реверберацией [19].

Помимо акустических деградаций, вызванных шумами и комнатой, встречаются и порчи сигналов от передачи аудио потока. Данные в реальном времени характеризуются высокой чувствительностью к задержкам. Ярким примером являются голосовые коммуникации: в телефонном разговоре задержка в одну-две секунды между высказыванием и ответом делает взаимодействие крайне неудобным. Аналогичные эффекты наблюдаются при трансляции новостей с зарубежных площадок, когда сигнал приходит с заметной задержкой [20]. В отличие от передачи файлов, повторная отправка потерянных пакетов нежелательна, так как это увеличивает задержку и нарушает последовательность информации. Кроме того, вариативность скорости поступления пакетов (джиттер) может вызвать непредсказуемое поведение системы. Таким образом, потеря пакетов приводит к пропаданию фрагментов

аудио, а джиттер – к вариативности времени прихода пакетов, что может приводить к ошибкам моделей распознавания речи.

Существуют разнообразные общедоступные наборы данных с зашумлённой или искажённой речью [18-19, 21-23]. Рассмотрим их основные характеристики в табл. 1.

Табл. 1. Обзор существующих речевых наборов данных с деградациями.
Table 1. Review of existing speech datasets with acoustic degradations.

Название набора	Описание	Языки
CHiME-5 [18]	Реальные разговоры, 20 домов, бытовой шум, запись на несколько микрофонов	71
VOiCES [19]	Корпус, созданный из чистой речи (LibriSpeech) с воспроизведением в помещениях с шумом и реверберацией; запись 12 микрофонами.	en
REVERB Challenge Dataset [21]	Корпус импульсных характеристик помещений. Включает как смоделированные (с искусственными АЧХ помещений), так и реальные записи речи в реверберационных помещениях. Предусмотрены записи с 1, 2 и 8 микрофонами.	en
DNS Challenge Dataset [22]	Набор для задач шумоподавления: содержит чистую речь, шумы, реверберацию и смесь голосов.	fr,ge,it,ru, sp,en
NISQA corpus [23]	Симулированные деградации речи: шумы, кодеки, потери пакетов, джиттер и др.	en

Несмотря на опубликованные наборы данных с деградациями, лишь небольшое число наборов содержит примеры контролируемых деградаций, вызванных нестабильным интернет-соединением, и позволяет исследовать влияние определенной степени деградаций на качество распознавания речи. Очень трудно понять, является ли какой-либо набор данных достаточно сложным для распознавания из-за качества аудио данных, а не из-за других причин (некачественная разметка и прочее).

2.3 Методы для работы в условиях акустических деградаций

Ключевым приёмом адаптации под акустические деградации является аугментация данных при обучении моделей. Одним из самых известных методов является SpecAugment [24]: к спектрограмме применяют искажение «Time Warp», затем маскируют произвольные блоки частот и времени (то есть намеренно удаляют части спектра). Это имитирует реальную потерю каналов или пропуск сегментов. Также активно используют добавление синтетических и естественных шумов с реверберацией – накладывают реальные фоны (уличный, бытовой, офисный шум) и искусственно сгенерированные импульсные отклики помещений. Например, вводят задержки согласно моделям помещений или свертки с реальными импульсными откликами (Room Impulse Responses), чтобы тренировать сети на «ревербированном» сигнале. Кроме того, применяют перестановку времени (speed perturbation) и другие техники (см. обзор [24]).

Для эмуляции потери пакетов в обучении искусственно «зануляют» фрагменты аудио. Авторы работы [25] показали, что при добавлении имитации единичных потерь до 10% в

обучающую выборку ASR-модель почти не теряет в точности. Однако при 15–20% потерь ошибки резко растут: лучшие стратегии аугментации дают качество распознавания на 7–10% выше, чем на чистых данных.

В ряде работ описывается апробация моделей распознавания речи для аудио с применением сетевых деградаций. В частности, авторы работы [26] предлагают методику аугментации речи сетевыми деградациями (потеря пакетов, джиттер) на основе известного набора данных зашумленной речи NOIZEUS. В результате генерируются около 192.5 часов речи, что позволяет изучать влияние искажений на распознавание речи в приложениях телефонии. Однако в работе не рассматривается сценарий кратковременной потери пакетов, которая характерна для реальных сетей и приводит к выпадению целых слов или словосочетаний. Такой сценарий особенно сложен для моделей, использующих контекстную информацию. Кроме того, созданные набор данных не опубликован.

Проблема восстановления потерянных пакетов (Packet Loss Concealment, PLC) остается актуальной, о чем свидетельствуют как специализированные исследования, например, работа [27], так и инициативы индустрии, такие как специализированный конкурс от Microsoft [28], направленный на исследование методов восстановления потерянных сетевых пакетов.

Другие подходы ориентированы на восстановление пропавших данных. Например, авторы [29] предлагают многослойную рекуррентную нейронную сеть, которая восстанавливает отсутствующие сегменты речи. Их модель способна восстанавливать до 1 секунды пропуска с высокой субъективной оценкой качества.

В архитектурном плане применяются также гибридные системы. Авторы работы [30] построили нейросеть, которая, будучи подключённой к модели распознавания речи с фиксированными весами, заполняет потерянные спектральные фреймы так, чтобы минимизировать ошибку. Кроме того, исследуются и трансформеры: например, модель Whisper (OpenAI) на 680 K часов многоязычных данных показала удивительную устойчивость к фоновым звукам [31]. Авторы этой работы демонстрируют, что модель Whisper лучше сохраняет качество распознавания при сильных шумовых зашумлениях, чем другие архитектуры распознавания речи.

3. Тестовый набор акустических деградаций, вызванных некачественным сетевым соединением

В отличие от классических акустических шумов (таких как реверберация, фоновая речь или уличный шум), деградации, возникающие из-за некачественного интернет-соединения, имеют природу сетевого происхождения. Они связаны не с физическим искажением звука, а с потерей, задержкой или фрагментацией цифровых пакетов в потоке данных. Такие дефекты проявляются нерегулярно и могут создавать эффекты «пропадания» частей речи, искажений тембра, обрыва слов или временной рассинхронизации между аудио и видео.

Особенность сетевых деградаций заключается в непредсказуемости и дискретности потерь: в отличие от аддитивных шумов, они нарушают структуру временной последовательности сигнала, что делает задачу реконструкции и последующего распознавания особенно сложной для ASR-систем. Потеря пакетов приводит к тому, что модель сталкивается с обрывами спектрограмм, изменением амплитудных соотношений и нарушением контекстных зависимостей, что снижает точность распознавания даже при неизменных условиях фонового шума.

В рамках данного исследования был подготовлен специальный набор данных, моделирующий влияние неустойчивого соединения на распознавание речи (рис. 1). Для симуляции использовались два режима порчи: потеря пакетов на протяжении всего аудио файла (полный режим – ***full***) и выборочная деградация для ***n*** последовательно идущих слов в тексте (частичный режим – ***partly***). Таким образом, такая комбинация позволяет создать

контролируемый характер порчи в наборах данных, что позволяет сравнивать модели распознавания речи и понимать границы их применимости в практических сценариях.

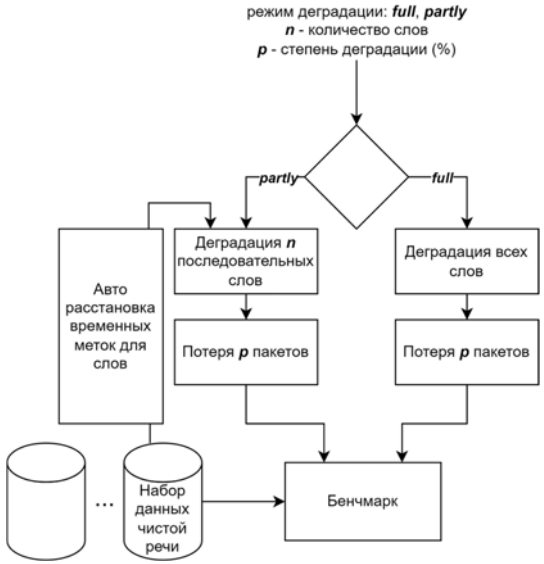

Рис. 1. Предлагаемый тестовый набор для эмуляции нестабильного интернет-соединения: схема генерации деградаций.
Fig. 1. The proposed benchmark for emulating an unstable Internet connection: a degradation generation scheme.

4. Автоматическая оценка деградаций для предсказания WER

В условиях непредсказуемых деградаций важно оперативно оценивать качество поступающего звукового потока, иметь возможность до запуска модели распознавания речи, понимать примерный уровень её качества. Такой подход позволяет не только экономить вычислительные ресурсы, но и своевременно диагностировать причины деградации, сигнализируя о проблеме в аудиопотоке.

Метрика Word Error Rate (WER) является стандартным показателем качества систем автоматического распознавания речи (ASR). Она измеряет расстояние между эталонном текстом и гипотезой, выданной моделью ASR, вычисляя минимальное количество операций на уровне слов – замен (S), вставок (I) и удалений (D), – необходимых для преобразования гипотезы в эталонный текст: $WER = (S + D + I) / N$ (где N – общее количество слов в эталонном тексте).

Для решения этой задачи мы используем метод предсказания метрики WER на основе анализа акустического сигнала без использования транскрипции [32, 33]. Предлагаемый нами подход основан на совокупности низкоуровневых и спектральных признаков, а также методов оценки качества при помощи автоматических метрик (рис. 2).

Формирование признаков для прогноза WER включает в себя оценки уверенности ASR-модели (метрики уверенности, извлечённые из логарифмических вероятностей на уровне токенов, такие как энтропия, коэффициент Gini и другие) и признаки качества речи (WADA-SNR [34], SpeechBrain SI-SNR Estimator [35], NISQA [23]), что позволяет учитывать влияние различных типов аудио деградаций. Объединённый вектор признаков подаётся на вход модели регрессии, в качестве которой используется алгоритм CatBoost [36].

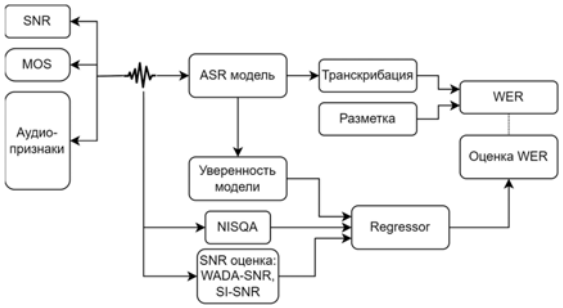

Рис. 2. Общая схема работы предсказания WER.
Fig. 2. The general scheme of the WER prediction.

5. Эксперименты

В данном разделе рассмотрим основные эксперименты по обучению и валидации моделей. Для экспериментов были выбраны 2 русскоязычных набора: LibriSpeech (ru) [37] и Fleurs (ru) [38]. Наборы содержат чистую речь, без акустических деградаций, поэтому подходят для контролируемой потери пакетов. Набор данных LibriSpeech (ru) основан на аудиокнигах, находящихся в открытом доступе в LibriVox. Набор данных Fleurs (ru) это речевая версия набора для машинного перевода FLoRes [39].

5.1 Оценка качества распознавания речи существующими моделями

Для экспериментов были выбраны следующие модели: Whisper-small, Whisper-large-v3-turbo, FastConformer и Qwen-3-Omni (табл. 2). В табл. 3 представлены результаты качества распознавания речи на наборе LibriSpeech (ru). Метрики качества посчитаны как для предобученной версий модели, без изменения весов, так и после обучения с помощью LoRA-адаптеров [40]. Эмпирически было выбрано значения ранга для LoRA-адатера, равное 256, что соответствует обучению 21.6% весов модели у версии whisper-turbo-v3 (использовались веса матриц Q, K, V и набор полносвязных слоев оригинальной архитектуры Whisper).

Табл. 2. Используемые модели распознавания речи.
Table 2. Used speech recognition models.

Модель	Количество параметров	Тип декодера
Whisper-small [13]	242M	Transformer
Whisper-turbo-v3 [13]	809M	Transformer
FastConformer [12]	120M	Гибрид (RNNT+CTC)
Qwen3-Omni [17]	30B	MoE Transformer

Тренировочная выборка была сформирована из обучающих выборок соответствующих наборов данных. Для этих выборок была применения процедура деградации, описанная в разделе 3 по потере пакетов (полный режим). Степень деградации ***p*** варьировалась от 0% до 90% с шагом 10%. Аналогичным образом применялась деградация для test выборок соответствующих наборов. У модели FastConformer обучались все веса (full-sft). Обучение всех моделей запускалось на 1 видеокарте Nvidia H100 80Gb. Среднеквадратичное отклонение у всех моделей < 0.2 по значениям WER, кроме случаев с Qwen3-Omni.

Табл. 3. Результаты распознавания с потерями на полной записи на наборе LibriSpeech (ru).
Table 3. Recognition results with losses on the full recording of the LibriSpeech dataset (ru).

Модель / Потеря пакетов		0%	10%	20%	30%	40%	50%	60%
Whisper-small	Pretrain	0.2296	0.26	0.3038	0.3752	0.479	0.6545	1.384
	LoRA-sft	0.1640	0.1766	0.1925	0.2132	0.2516	0.3136	0.4299
Whisper-turbo-v3	Pretrain	0.2037	0.2184	0.2350	0.2751	0.3291	0.4074	0.5309
	LoRA-sft	**0.1032**	**0.1099**	**0.1151**	**0.1322**	**0.1544**	**0.1937**	0.2775
Fast Conformer	Pretrain	0.4538	0.4552	0.4707	0.5159	0.6086	0.7495	0.8788
	Full-sft	0.1229	0.1231	0.1274	0.1384	0.1577	0.1993	**0.2749**
Qwen3-Omni	Pretrain	0.1109 ± 0.1251	0.1361 ± 0.1349	0.1722 ± 0.1470	0.2485 ± 0.1859	0.3785 ± 0.2240	0.8162 ± 9.2656	4.2209 ± 32.7964
Прогноз WER		0.079	0.103	0.122	0.158	0.226	0.342	0.547

По результатам анализа данных из табл. 2 можно сделать следующие выводы:

- Дообученная версия модели Whisper-turbo-v3-809M показывает наилучшие результаты практически среди всех уровней потерь (кроме 60%);
- Модель FastConformer-120M (Full-sft) показывает хорошую устойчивость. Её качество деградирует наиболее плавно с ростом потерь. На уровне 60% потерь её значение метрики WER (0.2749) почти такое же, как у лучшей модели Whisper (0.2775), при этом модель FastConformer имеет гораздо меньший размер (120M против 809M);
- Модель Qwen3-Omni (Pretrain) демонстрирует сильное падение качества при уровне потерь 60%, достигая значения WER, равного 4.2209. Такие значения говорят о наличии сильной галлюцинации модели, например, многократное повторение одной и той же фразы и полное несоответствие разметке;
- Прогнозирование метрики WER показывает реалистичные результаты, особенно на средних уровнях потерь (20–40%). На высоких уровнях потерь (60%) этот прогноз (0.547) оказывается пессимистичнее, чем показывают лучшие дообученные модели (~0.27-0.43).

В табл. 4 представлены результаты качества распознавания речи на наборе данных Fleurs (ru). Настройка гиперпараметров и модели обучения совпадают с описанной выше схемой. По результатам экспериментов можно сделать следующие выводы:

- Модель Qwen3-Omni показывает наилучший результат среди всех моделей на исходном наборе данных, однако сильно теряет в качестве при потерях пакетов выше 40%, достигая максимума значения WER в 2.0846. Такие экстремальные значения WER и высокое среднеквадратичное отклонение выражаются в наличии галлюцинаций моделей (подробнее в табл. 5);

Табл. 4. Результаты распознавания с потерями на полной записи на наборе Fleurs (ru).
Table 4. Recognition results with losses on the full recording of the Fleurs (ru).

Модель / Потеря пакетов		0%	10%	20%	30%	40%	50%	60%
Whisper-small	Pretrain	0.1593	0.1731	0.1957	0.2228	0.3016	0.5112	0.9449
	LoRA-sft	0.1502	0.1695	0.1835	0.2065	0.2527	0.3208	0.4450
Whisper-turbo-v3	Pretrain	0.2021	0.2230	0.2488	0.2952	0.3580	0.4590	0.6229
	LoRA-sft	0.1266	0.1338	0.1426	0.1498	0.1756	0.2221	0.3120
Fast Conformer	Pretrain	0.3655	0.3744	0.3868	0.4206	0.4896	0.6153	0.7733
	Full-sft	0.1516	0.1543	0.1585	0.1689	0.1839	**0.2155**	**0.2763**
Qwen3-Omni	Pretrain	**0.0392 ± 0.0625**	**0.0429 ± 0.0648**	**0.0516 ± 0.0731**	**0.0671 ± 0.0834**	**0.1227 ± 0.1264**	0.3688 ± 2.5370	2.0846 ± 14.5095
Прогноз WER		0.054	0.059	0.077	0.115	0.188	0.317	0.527

Табл. 5. Примеры галлюцинаций Qwen3-Omni на полной записи на наборе данных LibriSpeech (ru) 60% потери пакетов.
Table 5. Examples of Qwen3-Omni hallucinations on the full recording on the LibriSpeech dataset (ru) 60% packet loss.

Гипотеза модели	Правильная запись	Комментарий
Вот в этой чаше, в этой чаше, в этой чаше, в этой чаше, в этой чаше, в этой чаше … (повторяется более 100 раз)	отбойное течение течение создаваемое волнами которые разбиваются о берег или иногда об риф или что-то подобное	WER=256.0 из-за повторений словосочетания в гипотезе
Минск является административным центром Минской области.	не оскверняйте это место нанося или выцарапывая граффити на объекты вокруг	WER=1.0 гипотеза связная, но ни одно слово не совпадает
Для вас, пушкин, поэты, касается вас от них, времена, нувшему птице, в часы досуга золотых, чтоб в тайны подлинные рукой веры я впал.	для вас души моей царицы красавицы для вас одних времен минувших небылицы в часы досугов золотых под шепот старины болтливой рукою верной я писал	WER=0.75 есть единичные попадания по словам, но в общем контексте распознавание неверное

- Качество на наборе Fleurs у модели Qwen3-Omni лучше по сравнению с набором LibriSpeech. Набор Fleurs лучше соответствует домену тренировки модели Qwen3-Omni, чем набор LibriSpeech, поскольку в наборе LibriSpeech встречаются сказки русских писателей, и языковой стиль отличается от публицистического общего домена Fleurs;
- Модель FastConformer-120M (Full-sft) показывает наилучшую устойчивость к росту потерь. Наблюдается плавный рост метрики WER от 0.1516 при 0% потери пакетов до 0.2763 при 60%.

В табл. 6 и табл. 7 приведены результаты экспериментов на контролируемой частичной потере слов (см. раздел 3, режим ***partly***). В этом режиме случайным образом выбираются ***n*** идущих подряд слов, и к ним применяется потеря пакетов степенью деградации ***p***. Было эмпирически установлено, что на порче ограниченной последовательности слов имеет смысл рассматривать 2 сценария, как наиболее различимые на слух и по метрике WER: 40% и 60%.

Табл. 6. Результаты распознавания для потери на ограниченной последовательности слов (частичный режим) на наборе данных LibriSpeech (ru).
Table 6. Recognition results for the loss on a constrained word sequence (partly mode) on the LibriSpeech dataset (ru).

Модель / Потеря пакетов		40% 1 слово	40% 2 слова	40% 3 слова	60% 1 слово	60% 2 слова	60% 3 слова
Whisper-turbo-v3	Pretrain	0.2379	0.2500	0.2037	0.2341	0.2624	0.2956
	LoRA-sft	0.1940	0.2097	0.1521	0.1949	0.2242	0.2614
Fast Conformer	Pretrain	0.4631	0.4711	0.4538	0.4640	0.4854	0.5103
	Full-sft	0.1304	**0.1356**	**0.1229**	**0.1354**	**0.1523**	**0.1674**
Qwen3-Omni	Pretrain	**0.1241 ± 0.1321**	0.1468 ± 0.1473	0.1682 ± 0.1605	0.1487 ± 0.1433	0.2063 ± 0.1664	0.2671 ± 0.1930
Прогноз WER		0.097	0.111	0.123	0.111	0.144	0.177

Табл. 7. Результаты распознавания на ограниченной последовательности слов (частичный режим) на наборе данных Fleurs (ru).
Table 7. Recognition results for the loss on a constrained word sequence (partly mode) on the Fleurs dataset (ru).

Модель / Потеря пакетов		40% 1 слово	40% 2 слова	40% 3 слова	60% 1 слово	60% 2 слова	60% 3 слова
Whisper-turbo-v3	Pretrain	0.2215	0.2299	0.2021	0.2165	0.2366	0.2641
	LoRA-sft	0.1838	0.1938	0.1674	0.1842	0.2023	0.2247
Fast Conformer	Pretrain	0.3727	0.3779	0.3655	0.3747	0.3928	0.4088
	Full-sft	0.1559	0.1596	0.1516	0.1589	0.1689	0.1752
Qwen3-Omni	Pretrain	**0.0433 ± 0.0673**	**0.0458 ± 0.0693**	**0.0526 ± 0.0796**	**0.0500 ± 0.0723**	**0.0642 ± 0.0851**	**0.0909 ± 0.1025**
Прогноз WER		0.059	0.063	0.072	0.061	0.086	0.109

По результатам сравнения моделей в частичном режиме можно сделать следующие выводы:

- На всех моделях виден монотонный рост WER при увеличении теряемых слов на одном уровне потери пакетов. Однако, для набора LibriSpeech рост WER более выражен для всех моделей, чем у набора Fleurs в одинаковых режимах;
- Модель FastConformer все также продолжает демонстрировать наилучшую устойчивость: значение метрики WER растет с 0.1354 до 0.1674 (при потере пакетов 60%) на наборе LibriSpeech и с 0.1589 до 0.1752 (при такой же потере пакетов в 60%) на наборе Fleurs;
- При небольших деградациях относительно низкие значения предсказанной метрике WER говорят о том, что на этом домене можно обучить модель до хороших значений. Предобученные (pretrain) и дообученные (sft) модели это подтверждают;
- Большая разница в качестве у модели Qwen3-Omni (в 3 раза для режима с потерей 60% пакетов на одном слове) между двумя наборами указывает на сильную зависимость модели от текстового домена данных;
- Модель Qwen3-Omni стабильно лучше справляется с режимом частичной деградации ***partly***, при этом сильно падает в качестве по сравнению с другими моделями при полном режиме ***full***. Это связано с более мощным декодером данной модели по сравнению с другими (подробнее анализ результатов распознавания речи в табл. 8);
- Фактические значения метрики WER хуже, чем прогнозированные, для большинства моделей на наборах данных LibriSpeech и Fleurs.

Табл. 8. Примеры распознаваний Qwen3-Omni на ограниченной последовательности слов (частичный режим) на наборе данных Fleurs (ru).
Table 8. Examples of Qwen3-Omni recognition based on a limited sequence of words (partially mode) on the Fleurs dataset.

Whisper-turbo-v3	Fast Conformer	Qwen3-Omni
коммутации в клетках зародышевой линии могут передаваться далее детям в то время коммутации гдето еще могут привести к гибели клеток или раку	коммутации в клетках зародышевой линии могут передаваться далее детям в то время коммутации гдето еще могут привести к гибели клеток или раку	только мутации в клетках зародышевой линии могут передаваться далее детям в то время как мутации гдето еще могут привести к гибели клеток или раку
людям с этих времен были известны основные химические элементы такие как золото серебро и медь поскольку их все можно встретить в природе в чистом виде и относительно легко добывать с помощью примитивных орудий	людям с этих времен были известны основные химические элементы такие как золото серебро и медь поскольку их все можно встретить в природе в чистом виде и относительно легко добывать с помощью примитивных орудий	людям с древних времен были известны основные химические элементы такие как золото серебро и медь поскольку их все можно встретить в природе в чистом виде и относительно легко добывать с помощью примитивных орудий
хьюн ушел в остатку и в кабинете министров его заменит член парламента…	Хьюн ушел в остатку и в кабинете министров его заменит член парламента …	Хьюн ушел в отставку, и в кабинете министров его заменит член парламента…

5.2 Автоматическое предсказание метрики WER

Рассмотрим результаты экспериментов по автоматическому предсказанию WER (Раздел 4) на наборах LibriSpeech (ru) и Fleurs (ru). В первом (полном) режиме рассмотрим полную потерю пакетов на всем аудио файле. Во втором режиме деградация применяется случайно на ***n*** (1, 2, 3) последовательных слов в предложении – частичный режим. В табл. 9 представлены результаты по качеству автоматического предсказания метрики WER для трёх моделей на наборе Fleurs (ru).

Табл. 9. Качество прогнозирования WER на наборах данных Fleurs и LibriSpeech (ru).
Table 9. The quality of WER forecasting on the Fleurs and LibriSpeech datasets (ru).

Модель	Размер	Fleurs			Librispeech		
		RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
Whisper-turbo-large-v3	809M	**0.0768**	**0.9827**	0.9145	**0.0633**	**0.9891**	0.9424
Whisper-small	244M	0.1077	0.9527	**0.9399**	0.0752	0.9817	**0.9545**
Fast Conformer	120M	0.0866	0.9682	0.8942	0.0758	0.9828	0.9340

Результаты экспериментов проиллюстрированы на рис. 3, 4 и 5. Можно сделать несколько выводов:

- Предсказанная метрика WER коррелирует с долей потерянных пакетов. Как показано на рис. 3 и 4, в режиме полной деградации аудио-сигнала предсказанное значение метрики WER растет монотонно. Значения WER относительно небольшие (<0.2) при потерях до 30–40%, но резко растут начиная с 50% потерь, достигая медианного значения 1.0 при 90% потерь на обоих наборах;
- Увеличивается не только медианное значение метрики WER, но и дисперсия предсказаний. На рис. 3 и 4 разброс становится значительно шире при увеличении потерь до 70%. Это говорит о том, что при сильной деградации предсказания модели становятся менее стабильными;

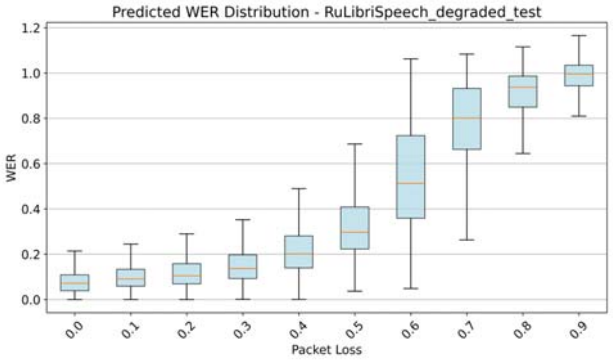

Рис. 3. Распределение предсказания WER на наборе данных LibriSpeech (ru) – full режим.
Fig. 3. Distribution of the WER prediction on LibriSpeech dataset (ru) – full mode.

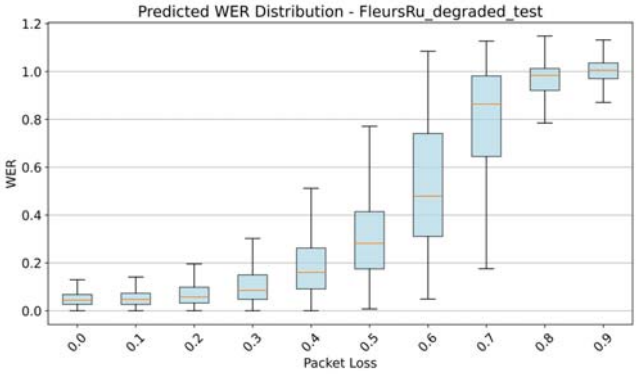

Рис. 4. Распределение предсказания WER на наборе данных Fleurs (ru) – full режим.
Fig. 4. Distribution of the WER prediction on Fleurs dataset (ru) – full mode.

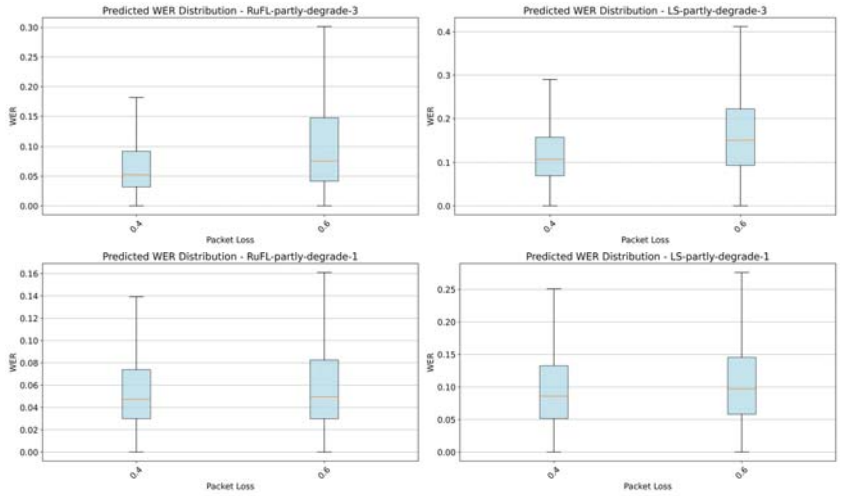

Рис. 5. Распределение предсказания WER на наборах данных Fleurs (ru) и LibriSpeech (ru) – частичный режим.
Fig. 5. Distribution of WER predictions on the Fleurs (ru) and LibriSpeech datasets (ru) – partly mode.

- Несмотря на потерю стабильности при потерях> 50% для значений WER до 0.4 это все равно хороший инструмент по прогнозированию на практике;
- Графики показывают, что при одинаковой доле потерь деградация большего числа слов оказывает более сильное влияние на прогноз WER. Так, на рис. 5 медианное значение WER при 40% потерь на трех словах выше, чем при 40% потерь на одном слове. Дисперсия с ростом деградации растет, как и на полном режиме ***full***;
- Однако прогнозирование WER не имеет потокового режима распознавания, поэтому совокупно показывает низкие значения на режиме потери на ограниченной последовательности слов (***partly***) по сравнению с потерей слов на всей записи (***full***).

6. Заключение

В рамках проведенного исследования исследовалась задача по оценке современных систем автоматического распознавания речи (ASR) к акустическим искажениям, характерным для нестабильного интернет-соединения:

- Был разработан и опубликован специализированный тестовый набор данных русскоязычной речи, ключевой особенностью которого является репрезентативный набор данных с контролируемыми и воспроизводимыми типами деградации аудио сигнала;
- На предложенном тестовом наборе была проведена сравнительная апробация современных подходов к распознаванию речи. Модель Qwen3-Omni стабильно лучше справляется с частичным режимом деградации, при этом сильно падает в качестве по сравнению с другими моделями при полном режиме. Это связано с хорошим контекстом Qwen3-Omni на стадии декодирования для генеративного предсказания следующих токенов по сравнению с другими моделями;
- Обученная версия модели FastConformer продемонстрировала наилучшую устойчивость для всех режимов тестового набора, например, для частичного режима: значение метрики WER находится в диапазоне от 0.1354 до 0.1674 (при потере пакетов 60%) на наборе данных LibriSpeech, и в диапазоне от 0.1589 до 0.1752 (при потере пакетов 60%) на наборе данных Fleurs;
- Для количественной оценки степени искажений был апробирован метод прогнозирования WER без использования сравнения с эталонным распознаванием речи, основанный на анализе совокупности акустических характеристик сигнала и автоматических нейросетевых метрик.

Список литературы / References

[1]. Feng Y., Devillers L. End-to-end continuous speech emotion recognition in real-life customer service call center conversations. B: 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2023.

[2]. Norda M., et al. Evaluating the efficiency of voice control as human machine interface in production. IEEE Transactions on Automation Science and Engineering, vol. 21, no. 3, 2023, pp. 4817-4828.

[3]. Venkatraman S., Overmars A., Thong M. Smart home automation – use cases of a secure and integrated voice-control system. Systems, vol. 9, no. 4, 2021: 77.

[4]. Russell S. O'Connor, et al. What automatic speech recognition can and cannot do for conversational speech transcription. Research Methods in Applied Linguistics, vol. 3, no. 3, 2024: 100163.

[5]. Pearsell S. M., Niebuhr O. Lost in the Noise: Evaluating ASR Performance in Industrial and Environment Noise. B: 2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS). IEEE, 2025.

[6]. Shah M. A., Raj B. Revisiting Acoustic Features for Robust ASR. B: ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[7]. Benesty J., et al. (ред.) Springer Handbook of Speech Processing. Berlin: Springer, 2008, vol. 1.

[8]. Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented Transformer for Speech Recognition. arXiv preprint arXiv:2005.08100, 2020. Доступно по ссылке: https://arxiv.org/abs/2005.08100, accessed 25.12.2025.

[9]. Yao Z., Guo L., Yang X., Kang W., Kuang F., Yang Y., Jin Z., Lin L., Povey D. Zipformer: A faster and better encoder for automatic speech recognition. arXiv preprint arXiv:2310.11230, 2023. Доступно по ссылке: https://arxiv.org/abs/2310.11230, accessed 25.12.2025.

[10]. Kim S., Gholami A., Shaw A., Lee N., Mangalam K., Malik J., Mahoney M. W., Keutzer K. Squeezeformer: An efficient transformer for automatic speech recognition. Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 9361-9373.

[11]. Peng Y., Dalmia S., Lane I., Watanabe S. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2022, pp. 17627-17643.

[12]. Rekesh D., et al. Fast Conformer with linearly scalable attention for efficient speech recognition. B: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023.

[13]. Radford A., et al. Robust speech recognition via large-scale weak supervision. B: Proceedings of ICML, PMLR, 2023.

[14]. Barański M., et al. Investigation of whisper ASR hallucinations induced by non-speech audio. ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.

[15]. Fang Q., et al. Llama-omni: Seamless speech interaction with large language models. arXiv preprint arXiv:2409.06666, 2024. Доступно по ссылке: https://arxiv.org/abs/2409.06666, accessed 25.12.2025.

[16]. Rubenstein P. K., et al. Audiopalm: A large language model that can speak and listen. arXiv preprint arXiv:2306.12925, 2023. Доступно по ссылке: https://arxiv.org/abs/2306.12925, accessed 25.12.2025.

[17]. Xu J., et al. Qwen3-omni technical report. arXiv preprint arXiv:2509.17765, 2025. Доступно по ссылке: https://arxiv.org/abs/2509.17765, accessed 25.12.2025.

[18]. Barker J., Watanabe S., Vincent E., Trmal J. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. B: Proc. Interspeech 2018, 2018, pp. 1561–1568. DOI: 10.21437/Interspeech.2018-1768.

[19]. Richey C., Barrios M. A., Armstrong Z., Bartels C., Franco H., Graciarena M., Lawson A., Nandwana M. K., Stauffer A., van Hout J., Gamble P., Hetherly J., Stephenson C., Ni K. Voices Obscured in Complex Environmental Settings (VOiCES) Corpus. B: Proc. Interspeech 2018, 2018, pp. 1566–1570. DOI: 10.21437/Interspeech.2018-1454.

[20]. Hartpence B. Packet Guide to Voice over IP: A system administrator's guide to VoIP technologies. O'Reilly Media, Inc., 2013.

[21]. Kinoshita K., Delcroix M., et al. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. EURASIP Journal on Advances in Signal Processing, 2016, Article 30. DOI: 10.1186/s13634-016-0306-6.

[22]. Reddy Ch. K. A., Gopal V., Cutler R., Beyrami E., Cheng R., Dubey H., Matusevych S., Aichner R., Aazami A., Braun S., Rana P., Srinivasan S., Gehrke J. The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. B: Proc. INTERSPEECH 2020, 2020, pp. 1036-1040.

[23]. Mittag G., Naderi B., Chehadi A., Möller S. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. B: Proc. Interspeech 2021, 2021, pp. 2127–2131. DOI: 10.21437/Interspeech.2021-299.

[24]. Park D. S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E. D., Le Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019. Доступно по ссылке: https://arxiv.org/abs/1904.08779, accessed 25.12.2025.

[25]. Fernández-Gallego M. P., Toledano D. T. A study of data augmentation for increased ASR robustness against packet losses. IberSPEECH, 2021.

[26]. Kumalija E., Elhard J., Nakamoto Y. Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech. Frontiers in Signal Processing, 2022, vol. 2, Article ID 999457. DOI: 10.3389/frsip.2022.999457.

[27]. Li, N., Zheng, X., Zhang, C., Guo, L., Yu, B. (2022) End-to-End Multi-Loss Training for Low Delay Packet Loss Concealment. Proc. Interspeech 2022, 585-589, DOI: 10.21437/Interspeech.2022-11439.

[28]. L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "Interspeech 2022 audio deep packet loss concealment challenge," in INTERSPEECH 2022 - 23rd Annual Conference of the International Speech Communication Association, 2022.

[29]. Shi H., Shi X., Dogan S. Speech inpainting based on multi-layer long short-term memory networks. Future Internet, vol. 16, no. 2, 2024: 63.

[30]. Dissen Y., et al. Enhanced ASR robustness to packet loss with a front-end adaptation network. arXiv preprint arXiv:2406.18928, 2024. Доступно по ссылке: https://arxiv.org/abs/2406.18928, accessed 25.12.2025.

[31]. Gong Y., et al. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers. arXiv preprint arXiv:2307.03183, 2023. Доступно по ссылке: https://arxiv.org/abs/2307.03183, accessed 25.12.2025.

[32]. Ali, Ahmed, and Steve Renals. "Word error rate estimation for speech recognition: e-WER." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics (ACL), 2018.

[33]. Polevoi A., Kragin A., Loukachevitch N. Ground Truth-Free WER Prediction for ASR via Audio Quality and Model Confidence Features. International Conference on Speech and Computer. Cham: Springer Nature Switzerland, 2025.
[34]. Kim C., Stern R. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. B: INTERSPEECH 2008, 2008, pp. 2598–2601. DOI: 10.21437/Interspeech.2008-644.
[35]. Subakan C., Ravanelli M., Cornell S., Grondin F. REAL-M: Towards Speech Separation on Real Mixtures. arXiv preprint arXiv:2110.10812, 2021. Доступно по ссылке: https://arXiv.org/abs/2110.10812, accessed 25.12.2025.
[36]. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516, 2019. Доступно по ссылке: https://arXiv.org/abs/1706.09516, accessed 25.12.2025.
[37]. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: an ASR corpus based on public domain audio books. B: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane (Australia), 2015, pp. 5206-5210. DOI: 10.1109/ICASSP.2015.7178964.
[38]. Conneau A., et al. Fleurs: Few-shot learning evaluation of universal representations of speech. 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023.
[39]. Goyal N., et al. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, vol. 10, 2022, pp. 522–538.
[40]. Hu E. J., et al. LoRA: Low-rank adaptation of large language models. Proceedings of ICLR, 2022.

Информация об авторах / Information about authors

Антон Вячеславович ПОЛЕВОЙ – аспирант Московского государственного университета им. М.В. Ломоносова на факультете вычислительной математики и кибернетики. Сфера научных интересов: цифровая обработка сигналов, машинное обучение для задач обработки звука, распознавания речи.

Anton Vyacheslavovich POLEVOI – postgraduate student at Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics. Research interests: digital signal processing, machine learning for audio processing tasks, automatic speech recognition.

Наталья Валентиновна ЛУКАШЕВИЧ – доктор технических наук, сотрудник научно-исследовательского вычислительного центра Московского государственного университета им. М.В. Ломоносова. Сфера научных интересов: автоматическая обработка языка, искусственный интеллект.

Natalia Valentinovna LOUKACHEVITCH – Dr. Sci. (Tech.), Lomonosov Moscow State University, Research Computing Center. Research interests: natural language processing, artificial intelligence.