



DOI: 10.15514/ISPRAS-2025-37(1)-1

Исследование влияния сетевых деградаций на модели распознавания речи

¹ И.И. Иванов, ORCID: 0000-0000-0000-0000 <ivanov@ispras.ru>

² П.П. Петров, ORCID: 0000-0000-0000-0000 <petrov@ispras.ru>

¹ Институт системного программирования им. В.П. Иванникова РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

² Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.

Аннотация. Несмотря на успехи моделей автоматического распознавания речи на различных датасетах и языках, применение моделей в повседневной жизни не позволяет использовать их в различных сценариях, например, таких как звонки с нестабильным сетевым соединением или телефонные каналы с помехами. В данной работе был разработан и представлен специализированный бенчмарк русскоязычной речи, ключевой особенностью которого является репрезентативный набор данных с контролируемыми деградациями сигнала, вызванными нестабильным интернет-соединением. На предложенном бенчмарке 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 бенчмарк речевых записей; аудио записи с нестабильным интернет-соединением.

Для цитирования: Иванов И.И., Петров П.П. Заголовок статьи. Труды ИСП РАН, том 37, вып. 1, 2025 г., стр. xx–xx. DOI: 10.15514/ISPRAS-2025-37(1)-1.

Благодарности: В этом блоке перечисляются организации, поддерживающие исследование, описанное в статье, гранты и т.д.

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

¹ I.I. Ivanov, ORCID: 0000-0000-0000-0000 <ivanov@ispras.ru>

² P.P. Petrov, ORCID: 0000-0000-0000-0000 <petrov@ispras.ru>

¹ Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

² Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, 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: Ivanov I.I., Petrov P.P. Article title. *Trudy ISP RAN/Proc. ISP RAS*, vol. 37, issue 1, 2025, pp. xx-xx (in Russian). DOI: 10.15514/ISPRAS-2025-37(1)-1.

Acknowledgements. Блок «Благодарности» на английском языке.

1. Введение

Развитие систем автоматического распознавания речи позволяет использовать модели для различных сценариев использования: автоматическая работа колл-центров [15], голосовое управление на предприятиях [16], в домах и автомобилях [17], протоколирование конференций и совещаний [10] 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 присутствии акустических деградаций [18], [19].

Под акустической деградацией будем понимать искажение исходного речевого сигнала, без изменение первоначального текста. Традиционно, выделяют несколько типов акустических искажений [20]:

- Реальные фоновые шумы различной природы: шум мотора, различных предприятий и заводов, городского транспорта и др.;
- Синтетические (сгенерированные) шумы: белый, розовый и др.;
- Реверберация внутри зданий и помещений (также возможна реверберация вне помещений, например, в горной местности, но встречается реже): высокие потолки, гулкие стены и др.;
- Дефекты устройств звукозаписи: клиппинг, 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 бенчмарки, 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 бенчмарка русской речи с контролируемым характером акустических деградаций из-за нестабильного интернет-соединения;
- Апробация современных подходов автоматического распознавания речи на предложенном бенчмарке;

- Автоматическая оценка деградаций при помощи подхода на основе совокупности акустических характеристик сигналов [8];
- 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 [1]. 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 адаптации офлайн 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, использующий U-Net-подобную структуру с многоуровневым downsampling слоем и повторным использованием весов внимания [2]. Авторы вводят новые функции нормализации (BiasNorm) и активации (Swoosh), что позволяет достичь сопоставимых или лучших результатов, чем у оригинального Conformer, при двукратном ускорении обучения;
- Squeezeformer [3] – тоже следует U-Net-подходу: центральные слои работают на пониженной частоте дискретизации, а блок обработки упрощается до вида стандартного Transformer. Это позволяет ускорить расчет при минимальной потере качества;
- Branchformer [4] – расширяет Conformer за счет параллельных ветвей самовнимания, что даёт модели возможность анализировать аудиосигнал разными способами одновременно;
- FastConformer [5] – 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 [9]). Авторы обучают модель сразу нескольким задачам, такие как: определение токена языка, выделение временных меток, а также возможности перевода на английский язык (если декодеру в служебных токенах подается метка «translate»). При этом они используют, в отличие от 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» [34] для русского языка.

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

- LLaMA-Omni [12] – объединяет предобученный аудио-энкодер, адаптер, LLM и потоковый декодер, благодаря чему может одновременно генерировать текст и голосовой ответ по устному запросу, не требуя явной промежуточной транскрипции;
- AudioPaLM [14] – объединяет текстовый LLM PaLM-2 и аудио модель AudioLM в единую мультимодальную архитектуру, достигающую высоких результатов в задачах ASR и перевода речи;
- Qwen3-Omni [13] – мультимодальная модель, демонстрирующая state-of-the-art на различных датасетах. Она обрабатывает как текст, так и аудио: поддерживает распознавание речи длиной до 40 минут, понимает свыше 100 языков и в целом превосходит закрытые аналоги (например, Gemini-2.5-Pro, GPT-4o-Transcribe) по точности распознавания.

Такие LLM-базированные модели 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) WER может изменяться в широких пределах. Например, в корпусе CHiME-5 записаны диалоги в 20 реальных домах с бытовыми шумами (кухня, кондиционер и т. д.) и разной акустикой [25]. Аналогично в VOiCES данные записаны в двух мебелированных комнатах с фоновыми шумами и реверберацией [26].

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 [30]. 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 [25, 26, 27, 28, 29]. Рассмотрим их основные характеристики в Таблице 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 [25]	Реальные разговоры, 20 домов, бытовой шум, запись на несколько микрофонов	en
VOiCES [26]	Корпус, созданный из чистой речи (LibriSpeech) с воспроизведением в помещениях с шумом и реверберацией; запись 12 микрофонами.	en

REVERB Challenge Dataset [27]	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 [28]	Набор для задач шумоподавления: содержит чистую речь, шумы, реверберацию и смесь голосов.	fr,ge,it,ru, sp,en
NISQA corpus [29]	Симулированные деградации речи: шумы, кодеки, потери пакетов, джиттер и др.	en

Несмотря на опубликованные датасеты с деградациями, довольно мало датасетов 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 пр.).

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 [11]: к спектрограмме применяют искажение «Time Warp», затем маскируют произвольные блоки частот и времени (т. е. 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 [11]).

Для эмуляции потери пакетов в обучении искусственно «зануляют» фрагменты аудио. Авторы показали [6], что при имитации единичных потерь до 10% ASR-модель с такими аугментированными данными почти не теряет в точности. Однако при 15–20% потерь ошибки резко растут: лучшие стратегии аугментации дают WER на 7–10% выше, чем на чистых данных [6].

Другие подходы ориентированы на восстановление пропавших данных. Например, RNN-inpainting: авторы [23] предлагают многослойную LSTM, 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 [24] построили U-Net-сеть, которая, подключённая к замороженной ASR-модели, заполняет потерянные спектральные фреймы так, чтобы минимизировать WER. Также исследуются и трансформеры: например, модель Whisper (OpenAI) на 680 К часов многоязычных данных показала удивительную устойчивость к фоновым звукам [7]. Авторы демонстрируют [7], что Whisper лучше сохраняет качество распознавания при сильных шумовых зашумлениях, чем другие архитектуры распознавания речи.

3. Бенчмарк 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.

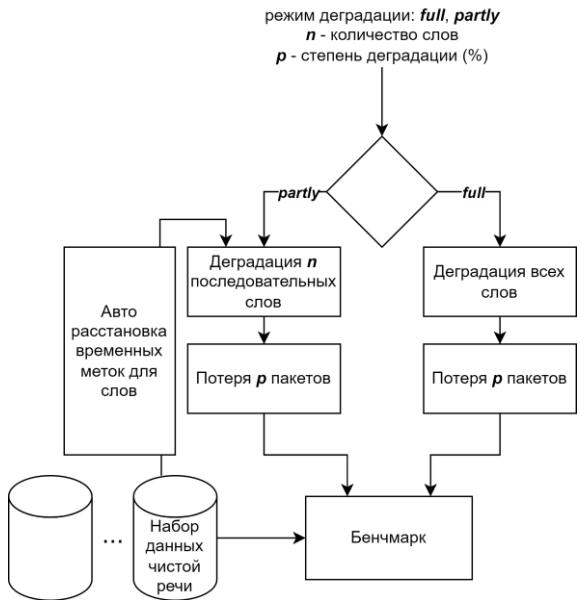


Рис. 1. Предлагаемый бенчмарк для эмуляции нестабильного интернет-соединения.

Fig. 1. The proposed benchmark for emulating an unstable Internet connection.

В рамках данного исследования был подготовлен специальный бенчмарк, моделирующий влияние неустойчивого соединения на распознавание речи (Рис. 1). Для симуляции использовались два режима порчи: потеря пакетов на протяжении всего аудио файла (режим *full*) и выборочная деградация для n последовательно идущих слов в тексте (режим *partly*). Таким образом, такая комбинация позволяет создать контролируемый характер порчи в датасетах, что позволяет сравнивать модели распознавания речи и понимать границы их применимости в практических сценариях.

4. Автоматическая оценка деградаций для предсказания WER

В условиях непредсказуемых деградаций важно оперативно оценивать качество поступающего звукового потока, иметь возможность до запуска модели с транскрипцией, понимать примерный уровень 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 эталонным 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 WER (Word Error Rate) на основе анализа акустического сигнала без привлечения транскрипции [8]. 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) и признаки качества речи (WADA-SNR [31], SpeechBrain SI-SNR Estimator [32], NISQA [29]), что позволяет учитывать влияние различных типов аудио деградаций. Объединённый вектор признаков подаётся на вход модели регрессии, в качестве которой используется алгоритм CatBoost [33].

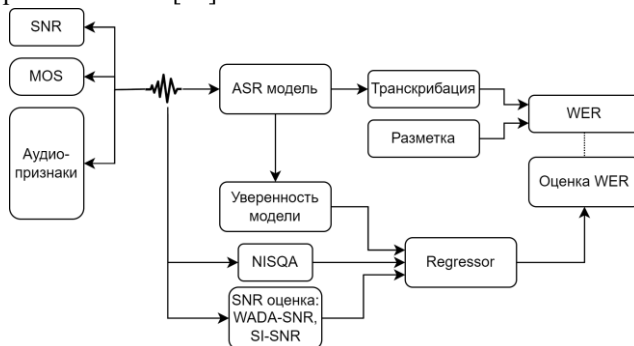


Рис. 2. Общая схема работы предсказания WER.

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) [21] и Fleurs (ru) [22]. Датасеты 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 [36].

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 (Таблица 2). В таблице 3 представлены результаты качества распознавания речи на датасете LibriSpeech (ru) [21]. Метрики качества посчитаны как для pretrain версий модели, без изменения весов, так и после обучения с помощью LoRa-адаптеров [35]. Поскольку эксперименты по полному обучению Whisper не позволили получить прироста в качестве, было принято решение обучать часть весов. Эмпирически было выбрано значения ранга для LoRa-адаптера, равное 256, что соответствует обучению 21.6% весов модели у версии whisper-large-v3-turbo (использовались веса модуля attention: матрицы Q, K, V и набор полносвязных слоев оригинальной архитектуры: fc1, fc2).

Табл. 2. Используемые модели распознавания речи.

Table 2. Used speech recognition models.

Модель	Количество параметров	Тип декодера
Whisper-small [9]	242M	Transformer
Whisper-turbo-v3 [9]	809M	Transformer
FastConformer [5]	120M	Гибрид (RNNT+CTC)
Qwen3-Omni [13]	30B	MoE Transformer

Тренировочная выборка была сформирована из обучающих выборок соответствующих датасетов. Для этих выборок была применена процедура деградации, описанная в разделе 3 по потере пакетов (режим *full*). 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.

Табл. 3. Результаты распознавания с потерями на полной записи на датасете 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	0.1361	0.1722	0.2485	0.3785	0.8162	4.2209
Прогноз WER		0.079	0.103	0.122	0.158	0.226	0.342	0.547

По результатам анализа данных из Таблицы 2 можно сделать следующие выводы:

- Дообученная версия модели Whisper-turbo-v3-809M показывает наилучшие результаты практически среди всех уровней потерь (кроме 60%);
- FastConformer-120M (Full-sft) показывает хорошую устойчивость. Её качество деградирует наиболее плавно с ростом потерь. На уровне 60% потерь её WER (0.2749) почти такой же, как у лучшей модели Whisper (0.2775), при этом FastConformer имеет гораздо меньший размер (120M против 809M);
- Модель Qwen3-Omni (Pretrain) демонстрирует сильное падение качества при уровне потерь 60%, достигая 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 WER показывает реалистичные результаты, особенно на средних уровнях потерь (20–40%). На высоких уровнях потерь (60%) его прогноз (0.547) оказывается пессимистичнее, чем показывают лучшие дообученные модели (~0.27-0.43).

Табл. 4. Результаты распознавания с потерями на полной записи на датасете 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	0.0429	0.0516	0.0671	0.1227	0.3688	2.0846
Прогноз WER		0.054	0.059	0.077	0.115	0.188	0.317	0.527

В таблице 4 представлены результаты качества распознавания речи на датасете Fleurs (ru) [22]. Настройка гиперпараметров и модели обучения совпадают с описанной выше схемой.

По результатам экспериментов можно сделать следующие выводы:

- Qwen3-Omni показывает наилучший результат среди всех моделей на исходном датасете, однако сильно теряет в качестве при потерях пакетов выше 40%, достигая максимума WER в 2.0846;
- Качество на Fleurs у Qwen3-Omni лучше по сравнению с LibriSpeech. Датасет Fleurs лучше соответствует домену тренировки Qwen3-Omni, чем LibriSpeech, поскольку в LibriSpeech встречаются сказки русских писателей и языковой стиль отличается от публицистического общего домена Fleurs;
- FastConformer-120M (Full-sft) показывает наилучшую устойчивость к росту потерь. Наблюдается плавный рост WER от 0.1516 при 0% потери пакетов до 0.2763 при 60%.

Также рассмотрим 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. 5. Результаты распознавания для потери на ограниченной последовательности слов (режим *partly*) на датасете LibriSpeech (ru).

Table 5. Recognition results for the loss on a constrained word sequence (*partly mode*) on the LibriSpeech dataset (ru).

Модель / Потеря пакетов		0%	40% 1 слово	40% 2 слова	40% 3 слова	60% 1 слово	60% 2 слова	60% 3 слова
Whisper-turbo-v3	Pretrain	0.2198	0.2379	0.2500	0.2037	0.2341	0.2624	0.2956
	Lora-Sft	0.1816	0.1940	0.2097	0.1521	0.1949	0.2242	0.2614
Fast Conformer	Pretrain	0.4575	0.4631	0.4711	0.4538	0.4640	0.4854	0.5103
	Full-sft	0.1258	0.1304	0.1356	0.1229	0.1354	0.1523	0.1674
Qwen3-Omni	Pretrain	0.1241	0.1468	0.1682	0.1109	0.1487	0.2063	0.2671
Прогноз WER		0.079	0.097	0.111	0.123	0.111	0.144	0.177

Табл. 6. Результаты распознавания на ограниченной последовательности слов (режим *partly*) на датасете Fleurs (ru).

Table 6. Recognition results for the loss on a constrained word sequence (*partly* mode) on the Fleurs dataset (ru).

Модель / Потеря пакетов		0%	40% 1 слово	40% 2 слова	40% 3 слова	60% 1 слово	60% 2 слова	60% 3 слова
Whisper-turbo-v3	Pretrain	0.2122	0.2215	0.2299	0.2021	0.2165	0.2366	0.2641
	Lora-Sft	0.1748	0.1838	0.1938	0.1674	0.1842	0.2023	0.2247
Fast Conformer	Pretrain	0.3695	0.3727	0.3779	0.3655	0.3747	0.3928	0.4088
	Full-sft	0.1530	0.1559	0.1596	0.1516	0.1589	0.1689	0.1752
Qwen3-Omni	Pretrain	0.0433	0.0458	0.0526	0.0392	0.0500	0.0642	0.0909
Прогноз WER		0.054	0.059	0.063	0.072	0.061	0.086	0.109

По результатам сравнения моделей в режиме *partly* 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 датасета Fleurs в одинаковых режимах;
- FastConformer все также продолжает демонстрировать наилучшую устойчивость: WER с 0.1354 до 0.1674 (при потере пакетов 60%) на датасете LibriSpeech и WER с 0.1589 до 0.1752 (при потере пакетов 60%) на наборе Fleurs;
- При небольших деградациях относительно низкие значения предсказанного WER говорят о том, что можно обучиться до хороших значений на этом домене. И pretrain и sft модели это подтверждают;
- Большая разница в качестве у модели Qwen3-Omni (в 3 раза для 1 слова 60%) между двумя наборами указывает на сильную зависимость модели от текстового домена данных;
- Модель Qwen3-Omni стабильно лучше справляется с *partly* режимом деградации, при этом сильно падает в качестве при *full* режиме по сравнению с другими моделями. Это связано с более мощным декодером для генеративного предсказания

следующих токенов по сравнению с другими моделями;

- Фактические значения WER хуже, чем прогнозируемые для большинства моделей на датасетах LibriSpeech и Fleurs.

5.2 Автоматическое предсказание WER

Рассмотрим результаты экспериментов по автоматическому предсказанию WER (Раздел 4) на датасетах LibriSpeech (ru) [21] и Fleurs (ru) [22]. В первом режиме **full** рассмотрим полную потерю пакетов на всем аудио файле. Во втором режиме деградация применяется случайно на n (1, 2, 3) последовательных слов в предложении – режим **partly**.

Результаты экспериментов представлены на рисунках 3, 4 и 5. Можно сделать несколько выводов:

- Предсказанный WER коррелирует с долей потерянных пакетов. Как показано на рисунках 3 и 4, в режиме полной деградации аудио предсказанный WER растет монотонно. Значения WER относительно небольшие (<0.2) при потерях до 30–40%, но резко растет начиная с 50% потерь, достигая медианного значения 1.0 при 90% потерь на обоих датасетах;
- Увеличивается не только медианный WER, но и дисперсия предсказаний. На рисунках 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;

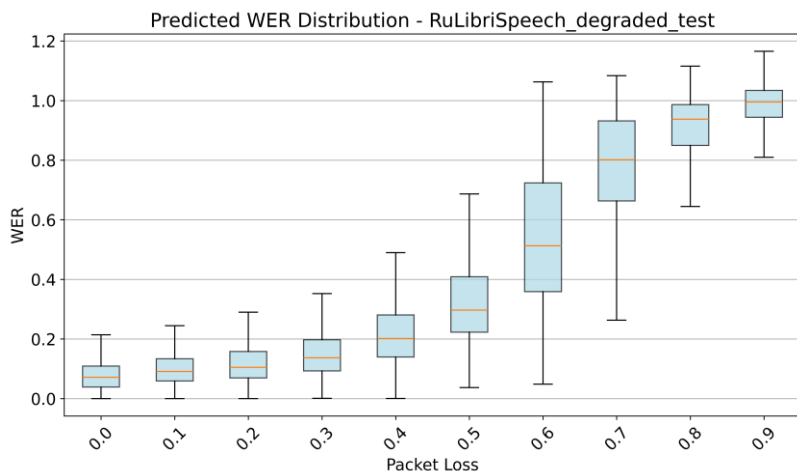


Рис. 3. Распределение предсказания WER на датасете LibriSpeech (ru) – full режим.

Fig. 3. Distribution of the WER prediction on LibriSpeech dataset (ru) – full mode.

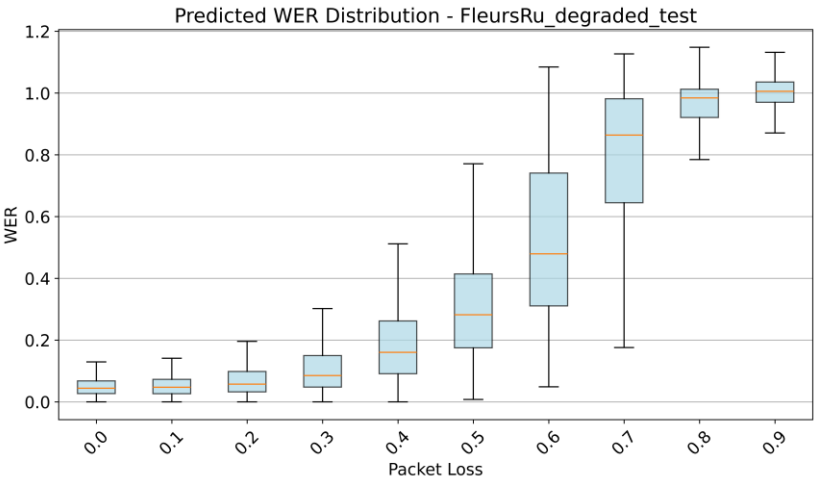


Рис. 4. Распределение предсказания WER на датасете Fleurs (ru) – full режим.
Fig. 4. Distribution of the WER prediction on Fleurs dataset (ru) – full mode.

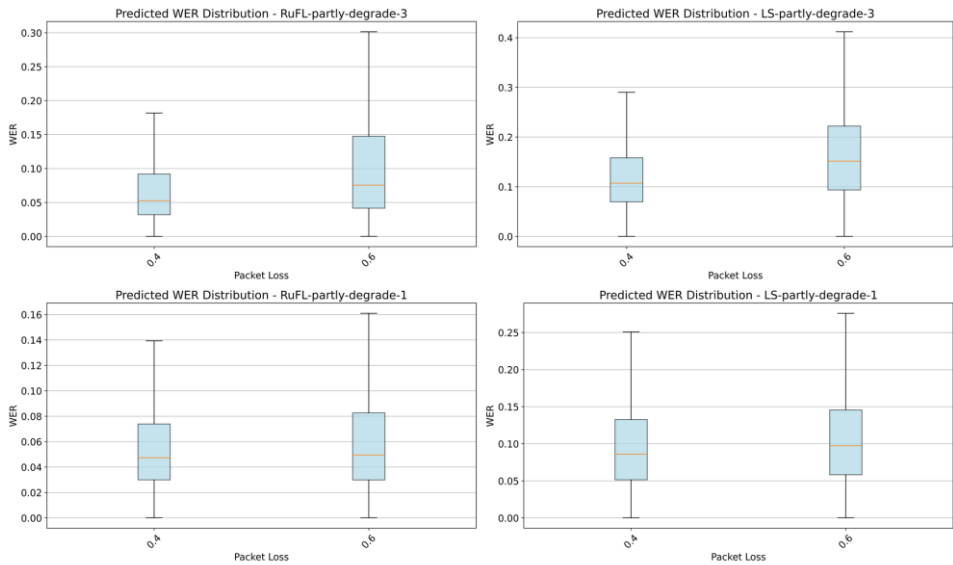


Рис. 5. Распределение предсказания WER на датасетах Fleurs (ru) и LibriSpeech (ru) – partly режим.
Fig. 5. Distribution of WER predictions on the Fleurs (ru) and LibriSpeech datasets (ru) – partly mode.

- 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 рисунке 5 медианный WER при 40% потерь на 3 словах выше, чем при 40% потерь на 1 слове. Дисперсия с ростом деградации растет, как и на **full** режиме;
- Однако прогнозирование WER не имеет потокового инференса, 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 бенчмарк русскоязычной речи, ключевой особенностью которого является репрезентативный набор данных с контролируемыми и воспроизводимыми типами деградации аудио сигнала;
- На предложенном бенчмарке была проведена сравнительная апробация современных подходов к распознаванию речи. Модель Qwen3-Omni стабильно лучше справляется с *partly* режимом деградации, при этом сильно падает в качестве при *full* режиме по сравнению с другими моделями. 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 бенчмарка, например, для режима *partly*: WER с 0.1354 до 0.1674 (при потере пакетов 60%) на датасете LibriSpeech и WER с 0.1589 до 0.1752 (при потере пакетов 60%) на наборе Fleurs;
- Для количественной оценки степени искажений был апробирован метод прогнозирования WER без использования разметки, основанный на анализе совокупности акустических характеристик сигнала и автоматических нейросетевых метрик.

Список литературы / References

- [1]. 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>.
- [2]. 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>.
- [3]. 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.
- [4]. 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.
- [5]. 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.
- [6]. Fernández-Gallego M. P., Toledano D. T. A study of data augmentation for increased ASR robustness against packet losses. IberSPEECH, 2021.
- [7]. 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>.
- [8]. Polevoi A., Kragin A., Loukachevitch N. Ground Truth-Free WER Prediction for ASR via Audio Quality and Model Confidence Features. In Speech and Computer (SPECOM), Cham: Springer, 2025.
- [9]. Radford A., et al. Robust speech recognition via large-scale weak supervision. B: Proceedings of ICML, PMLR, 2023.
- [10]. 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.
- [11]. 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>.
- [12]. 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>.

- [13]. Xu J., et al. Qwen3-omni technical report. arXiv preprint arXiv:2509.17765, 2025. Доступно по ссылке: <https://arxiv.org/abs/2509.17765>.
- [14]. 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>.
- [15]. 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.
- [16]. 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.
- [17]. 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.
- [18]. 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.
- [19]. 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.
- [20]. Benesty J., et al. (ред.) Springer Handbook of Speech Processing. Berlin: Springer, 2008. Том 1.
- [21]. 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.
- [22]. Conneau A., et al. Fleurs: Few-shot learning evaluation of universal representations of speech. 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023.
- [23]. 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.
- [24]. 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>.
- [25]. 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.
- [26]. 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.
- [27]. 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.
- [28]. Reddy Ch. K. A., Gopal V., Cutler R., Beyrami E., Cheng R., Dubey H., Matushevych 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.
- [29]. 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.
- [30]. Hartpence B. Packet Guide to Voice over IP: A system administrator’s guide to VoIP technologies. O’Reilly Media, Inc., 2013.
- [31]. 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.
- [32]. 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>.
- [33]. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulín A. CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516, 2019. Доступно по ссылке: <https://arxiv.org/abs/1706.09516>.
- [34]. 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.
- [35]. Hu E. J., et al. LoRA: Low-rank adaptation of large language models. Proceedings of ICLR, 2022.
- [36]. 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.