УДК 004.85

ПРЕДСКАЗАНИЕ КАЧЕСТВА АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

А. В. Полевой[0009-0000-2216-0468]

Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия

polevoianton@bk.ru

Аннотация

Предложен подход к прогнозированию показателя качества распознавания речи Word Error Rate (WER) на основе акустических характеристик сигнала и вычисления перплексии языковых моделей. Предлагаемый метод включает в себя создание разнообразных наборов аудиоданных путем применения различных типов акустических искажений к чистым речевым образцам на различных уровнях качества и разборчивости. В отличие от предыдущих работ, извлекается и анализируется полный набор речевых признаков: прогнозирование значения отношения сигнал/шум (signal-to-noise ratio, SNR), нейросетевые метрики качества звука (NISQA и др.), метрики уверенности модели распознавания речи, а также перплексия текста гипотезы ASR по языковой модели в качестве дополнительного признака для обучения единой модели прогнозирования WER.

Проведены эксперименты с использованием современных архитектур распознавания речи для демонстрации эффективности предлагаемого метода в прогнозировании WER в различных акустических условиях. Показано, что включение перплексии существенно повышает качество прогноза WER, в частности для данных, где акустические признаки слабо коррелируют с ошибками распознавания. Результаты применимы для автоматической оценки ожидаемого качества распознавания речи и фильтрации аудиовходов.

Ключевые слова: *прогнозирование WER, акустическая деградация при распознавании речи, перплексия, уверенность систем автоматического распознавания речи.*

ВВЕДЕНИЕ

Системы автоматического распознавания речи активно внедряются в различные приложения: анализ диалога с клиентами [1], мониторинг подозрительных событий [2], управление голосом на предприятиях [3] и в повседневной жизни. Особенно важна надёжность распознавания при внедрении моделей в критические области, которые характеризуются повышенным вниманием к контролю и надежности выдаваемых результатов (это авиация, медицина, автопилоты и др.). Ошибка в таких областях обходится очень дорого и использование моделей становится невозможным, даже если ошибка возникает редко [4].

В большинство работ, связанных с распознаванием речи, уделяется внимание сбору и замеру метрик на различных наборах данных с разными тематическими доменами и языками [5]. Однако недостаточно исследован вопрос предсказания оценки качества распознавания речи на выбранных датасетах, а также сложности конкретных примеров из набора данных для распознавания. Кроме того, традиционные методы измерения качества используют метрику WER, для которой важна эталонная выборка для оценки, но для применений в реальном времени это представляется невозможным.

Обычно модели по распознаванию речи ошибаются из-за присутствующих в записи акустических деградаций и недостаточного уровня качества непосредственно самой речи спикеров:

- акустические деградации: шум окружающей среды, микрофонные артефакты, реверберация и искажения сигнала – все это можно имитировать для проведения тестов;
- сложность речи: акценты, редкие слова и термины, относящиеся к конкретной предметной области, для оценки надежности которых требуются специальные наборы данных.

Сложность речи по записи оценить в реальном времени трудно, не имея разметки. Для оценки терминологии требуется сначала получить транскрибацию записи и провести ее последовательный анализ, например, при помощи различных средств автоматического анализа текстов. Более того, если транскрибация получилась некачественной, то для текстового анализа необходимо использовать ручную разметку, что является неприменимым для приложений в реальном времени. Что же касается акустических деградаций, их диагностика

и мониторинг возможны без применения текстовой разметки. Зависимость текущих акустических деградаций от уровня ошибок ASR-модели можно построить напрямую. Моделируя эту взаимосвязь, необходимо симулировать как общеизвестные (клипинг, реверберация различных комнат), так и встречающиеся в целевом приложении акустические деградации. Таким образом, настоящая работа посвящена следующим целям:

- построение взаимосвязи симулированных и контролируемых деградаций и ошибок ASR-модели;
- автоматическая оценка WER/CER в потоковом режиме для заранее известного набора ASR-моделей;
- автоматический анализ текстового качества полученной записи распознанной речи на основе перплексии, рассчитанной по большой языковой модели (подробнее в п. 3.2). Предполагается, что чем выше перплексия, тем хуже полученная запись соответствует обученному многообразию предложений языковой модели на естественном языке.

1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Характерные для ASR-моделей акустические деградации

Проведенные исследования в области устойчивой работы автоматического распознавания речи (Automatic Speech Recognition, ASR) в основном были сосредоточены на защите моделей от акустических искажений. По данным работы [6], эти подходы можно разделить на две категории: основанные на моделях и основанные на признаках.

Подходы, основанные на модели, включают адаптацию предварительно обученных моделей [7, 8], предварительную обработку аудио с использованием техник шумоподавления [9, 10] и обучение непосредственно на зашумленных данных [11]. Эти стратегии обычно требуют доступа к репрезентативным зашумленным данным и наиболее эффективны, когда среда развертывания и характеристики шума известны заранее.

Подходы, основанные на признаках, сосредоточены на разработке представлений речи, инвариантных к шуму. К ним относятся признаки [12–14], взятые из биологии, и методы обработки сигналов, предназначенные для извлечения

релевантных компонентов речи при одновременной фильтрации нерелевантных элементов сигнала.

В последние годы было продемонстрировано, что крупномасштабное обучение на разнообразных акустических сценариях может создавать по своей сути устойчиво работающие системы ASR без применения специализированных методов. В частности, в работе [15] авторы показали, что модели, подобные Whisper, достигают существенной устойчивости благодаря обучению в разнообразных аудиоусловиях, в том числе с полуразмеченными данными.

Как правило, производительность новой системы ASR оценивается на нескольких стандартных наборах данных речи, которые по своей природе содержат различные уровни шума и качества сигнала. Сам по себе этот подход дает некоторое представление об устойчивости системы. Однако более прямой метод оценки устойчивой работы в различных условиях включает моделирование различных искажений сигнала и измерение частоты ошибок по словам (показатель WER) при разных уровнях отношения сигнал/шум (показатель SNR), как это продемонстрировано в многочисленных исследованиях по устойчивости ASR [15–18].

1.2. Метрики оценки качества моделей распознавания речи

Частота ошибок по словам (Word Error Rate, WER) является стандартной метрикой оценки для систем автоматического распознавания речи. Она измеряет расстояние между эталонной расшифровкой и гипотезой модели ASR путем вычисления минимального количества операций на уровне слов (замен, вставок и удалений), необходимых для преобразования гипотезы в эталон. Аналогично частота ошибок по символам (Character Error Rate, CER) измеряет расстояние на уровне символов (букв) и часто используется для языков с богатой морфологией или при оценке посимвольного распознавания. WER вычисляется по формуле

$$\mathrm{WER} = \frac{S+D+I}{N},$$

где S – количество замененных слов, D – количество удаленных слов, I – количество вставленных слов, N – общее количество слов в эталонной расшифровке. Более низкие значения WER указывают на лучшую производительность ASR, при этом 0.0 представляет собой идеальную расшифровку. WER может превышать

1.0 в случаях, когда количество ошибок вставки особенно велико (генерация текста на тишину и другие примеры).

1.3. Прогнозирование WER

Поскольку WER служит эффективным индикатором производительности ASR, основная мотивация для разработки методов прогнозирования WER заключается в оценке качества прогнозов ASR без наличия эталонных расшифровок, которые часто недоступны в реальных приложениях.

Было предложено несколько методов для прогнозирования WER без эталонных расшифровок. Например, подход eWER [19] позволяет обнаружить расхождения между двумя системами ASR (на уровнях слов и символов) и использует их для прогнозирования WER. eWER-2 [20] использует акустические, лексические и фонотактические признаки для той же задачи. eWER-3 [21] продемонстрировал, что их многоязычная модель превосходит предыдущие одноязычные методы прогнозирования WER (eWER2), достигая 9%-ного абсолютного увеличения коэффициента корреляции Пирсона, что показывает лучшую корреляцию между прогнозируемым и эталонным значениями WER. Общим недостатком этих методов является необходимость либо в двух отдельных моделях ASR, либо в дополнительных моделях для обработки текста и извлечения фонем, что увеличивает вычислительные требования и общую сложность.

В работе [22] авторы рассчитали просодические статистики (частоты, значения энергии, темп, паузы) и использовали модель машинного обучения для прогнозирования WER для записей телефонных диалогов. В [23] проведено исследование методов измерения качества речи, таких как POLQA [24], для прогнозирования WER. Эти результаты привели к созданию полиномиальных моделей для прогнозирования точности распознавания речи на основе инструментальных измерений при различных канальных искажениях в разных полосах пропускания.

Предлагаемый нами подход отличается несколькими ключевыми аспектами:

- WER прогнозируется на основе признаков качества аудио и при необходимости перплексии гипотезы по языковой модели. Данная мера показывает,

насколько распознанный текст естественен для языковой модели, высокие значения часто сопутствуют ошибочным гипотезам;

- проведены эксперименты с наиболее распространенными современными архитектурами ASR, а именно Whisper и FastConformer;
- подход основан на легковесной модели регрессора WER в сочетании с модульной архитектурой извлечения признаков. Эта конструкция обеспечивает гибкость, позволяя исследователям включать новые методы оценки качества речи по мере их появления;
- приведены сравнения важности признаков для содействия дальнейшим исследованиям в области прогнозирования WER и метрик качества речи.

2. ОПИСАНИЕ РЕШЕНИЯ ПО ПРОГНОЗИРОВАНИЮ WER

2.1. Данные

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.

Предыдущие исследования [25] опирались на классические синтетические шумы и известные шумовые паттерны. В рамках настоящего исследования был подготовлен специальный бенчмарк, моделирующий влияние неустойчивого соединения на распознавание речи. Для симуляции был использован режим потери пакетов на протяжении всего аудиофайла. Уровень потери пакетов изменялся от 0.0 (отсутствие потери пакетов, речь идеально различима на слух) до 0.9

(сильная деградация, речь неразличима на слух) с шагом 0.1. Такое разнообразие различных по сложности акустических деградаций позволило построить надежную модель прогнозирования уровня WER.

2.2. Прогнозирование WER

Для прогнозирования WER в рамках нстоящего исследования была разработана специализированная архитектура системы обработки данных и обучения модели (см. рис. 1). Основная цель системы – это извлечение информативных признаков из аудиозаписей и их транскриптов с последующим построением регрессионной модели, способной предсказывать значение WER для новых пар аудио – текст.

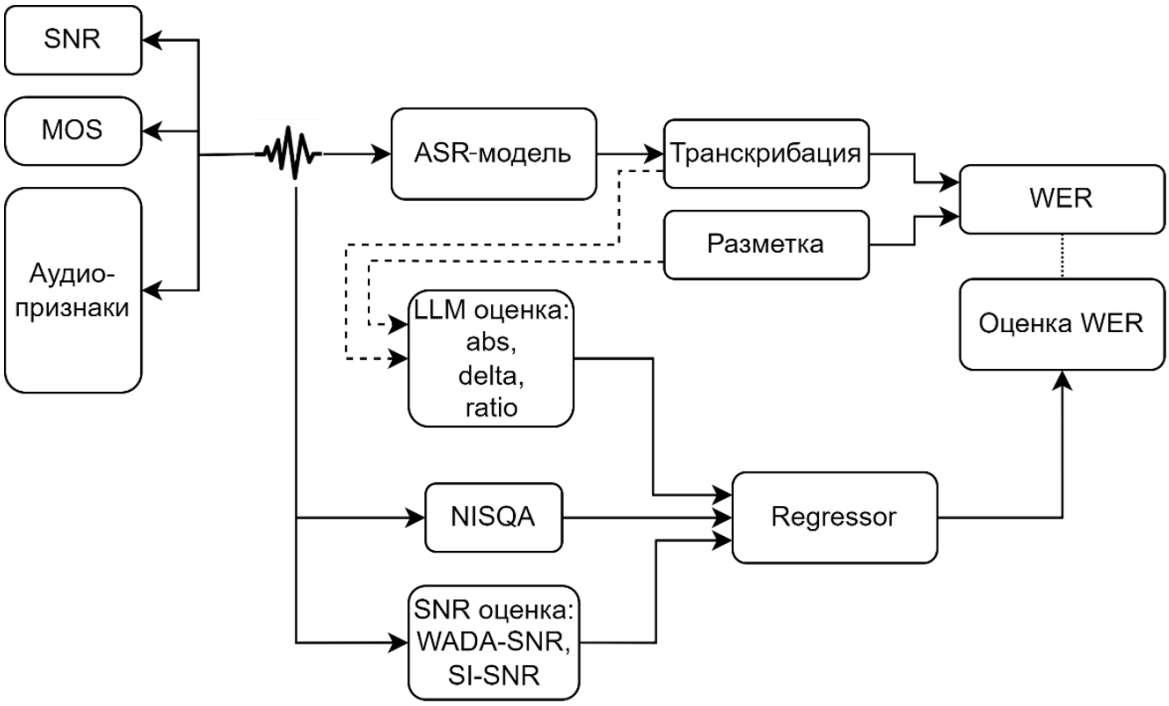

Рис. 1. Общая архитектура системы для прогнозирования WER.

Как видно из рис. 1, процесс прогнозирования WER начинается с загрузки исходных данных: аудиофайлов в формате WAV и соответствующих им текстовых транскриптов. Среди текстовых признаков в явном виде использована перплексия гипотезы ASR (см. п. 2.4): она характеризует степень соответствия распознанного текста типичным для языка конструкциям; при росте числа ошибок распознавания гипотеза часто содержит редкие или грамматически аномальные по-

следовательности, что приводит к повышению перплексии и позволяет использовать ее как предиктор WER наряду с акустическими признаками и признаками уверенности.

Были использованы следующие группы признаков:

1) аудиометрики – акустические характеристики сигнала, полученные через нейросетевые метрики;

2) признаки на основе перплексии: перплексия текста гипотезы ASR по языковой модели; перплексия текста, рассчитанная как разность или отношение между текстовой гипотезой и эталонной разметкой.

2.3. Аудиометрики

В контексте объективной оценки качества речи аудиометрики делятся на две основные группы: модели, оценивающие воспринимаемое качество, и объективные измерения соотношения сигнал – шум (signal to noise ratio, SNR).

К первой группе относится модель NISQA [26] (Non-Intrusive Speech Quality Assessment), которая на основе глубокого обучения предсказывает не только общую оценку качества (nisqa_mos), но и отдельные перцептивные аспекты: зашумленность (nisqa_noi), искажения тембра (nisqa_col), разрывы / цифровые потери (nisqa_disc) и громкость (nisqa_loud). На основе этих размерностей рассчитывается усредненный показатель (nisqa_average) для комплексной оценки деградации сигнала.

Вторую группу составляют модели SNR-меры того, насколько больше содержится полезного сигнала по сравнению с шумовой составляющей в исходной записи. Обычно SNR определяют как отношение энергий сигнала и шума и выражают в децибелах, $\mathrm{SNR}_{\mathrm{dB}} = 20\log_{10}\left(\frac{P_s}{P_n}\right)$. В этой группе выделяется метод WADA-SNR [27], основанный на анализе распределения амплитуд формы волны. Метрики из фреймфорка SpeechBrain (speechbrain_snr_a, speechbrain_snr_b) [28] реализуют вычисление отношения сигнал – шум (SI-SNR), которое оценивает искажения формы сигнала.

2.4. Вычисление перплексии

Перплексия последовательности слов $w_1, \dots, w_n$ по языковой модели определяется как экспонента средней отрицательной логарифмической вероятности:

$$\mathrm{PPL}(W) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log p(\, w_i \mid w_1, \dots, w_{i-1}\,)\right).$$

Чем выше перплексия, тем хуже последовательность согласуется с распределением, задаваемым языковой моделью; для ошибочных или неестественных гипотез ASR перплексия, как правило, возрастает. В настоящей работе перплексия гипотезы вычисляется с помощью предобученных языковых моделей (GPT-2 [29], Qwen [30], GigaChat3 [31], YandexGPT [32], Ruadapt [33]): по гипотезе получают логарифмы вероятностей токенов и по ним вычисляют значение PPL по приведенной ныше формуле.

2.5. Агрегация признаков и обучение модели

На этапе конструирования признаков параллельно обрабатываются оба типа данных (звук и текст). Акустические признаки и признаки уверенности агрегируются по временной оси; перплексия вычисляется для всей гипотезы целиком как разность (delta) или отношение (ratio) между текстовой гипотезой и эталонной разметкой. Полученные векторы конкатенируются и подаются регрессору на вход модели.

На втором этапе объединенный набор данных с целевой переменной WER разделяется на обучающую и тестовую выборки. На обучающих данных происходит настройка регрессора CatBoost [34], выбранного в силу эффективности его работы с категориальными признаками и устойчивости к переобучению. Итоговая модель оценивается на тестовой выборке, а ее прогнозы интерпретируются для анализа признаков, наиболее значимо влияющих на качество распознавания речи.

3. ЭКСПЕРИМЕНТЫ

Далее представлены результаты экспериментов по прогнозированию WER. Целью экспериментов была оценка эффективности различных стратегий

формирования признакового пространства для задачи регрессионного прогнозирования WER.

Для проведения экспериментов были использованы два набора данных:

- **LibriSpeech** [35] – корпус, состоящий из аудиокниг, записанных диктором;
- **Fleurs** [36] – набор данных для распознавания речи, речевая версия набора данных для машинного перевода FLoRes.

В качестве базовых моделей автоматического распознавания речи (ASR) были использованы Whisper-turbo-large-v3 (809M) и FastConformer (ru, 120M). Для каждой пары «аудио – текст» извлекались признаки: аудиометрики ('nisqa_noi', 'nisqa_col', 'nisqa_disc', 'nisqa_loud', 'nisqa_mos', 'nisqa_average', 'wada_snr', 'speechbrain_snr_a', 'speechbrain_snr_b'), перплексия текстовой гипотезы от модели распознавания речи и разность/отношение в перплексии между гипотезой и эталонной разметкой (delta_ppl, ratio_ppl). Перплексии вычислялись по формуле из п. 2.4 и при помощи языковых моделей из табл. 1.

Табл. 1. LLM, используемые для оценки перплексии текстовой гипотезы и эталонной разметки при подсчетах delta_ppl, ratio_ppl.

Модель в эксперименте	Полное название (HF)	Описание
Qwen2.5 [29]	Qwen/Qwen2.5-0.5B	0.49B
GPT-2 [30]	openai-community/gpt2-large	774M
GigaChat3 [31]	ai-sage/GigaChat3-10B-A1.8B	10B (1.8B активных параметров)
YandexGPT-5 [32]	yandex/YandexGPT-5-Lite-8B-instruct	8B
RuAdapt [33]	RefalMachine/RuadaptQwen3-8B-Hybrid	8B

Для оценки качества прогнозирования использовалась метрика RMSE (Root Mean Square Error) – среднеквадратичная ошибка (чем меньше, тем лучше).

Для оценки вклада перплексии текста гипотезы ASR в прогнозирование WER были проведены отдельные эксперименты на подмножествах набора Fleurs и LibriSpeech с двумя конфигурациями моделей: Whisper-turbo-large-v3 (809M) и FastConformer (ru, 120M). В качестве регрессоров были использованы

линейная регрессия (LR), Random Forest (RF) и CatBoost (CB). Для финальных результатов был выбран CatBoost как оптимальное решение. В табл. 2 представлены результаты, полученные по метрике RMSE.

Табл. 2. Прогнозирование WER по метрике RMSE: сравнение группы признаков.

Эксперимент	LLM	Fleurs Conformer	Fleurs Whisper	LS Conformer	LS Whisper
Аудиопризнаки	–	0.1976	0.1078	0.1891	0.2163
Аудиопризнаки + ppl гипотеза	Gpt2	0.1822	0.1018	0.1826	0.1780
Аудиопризнаки + delta_ppl, ratio_ppl	Gpt2	0.1544	**0.0730**	0.1400	**0.1329**
Аудиопризнаки + ppl гипотеза	Qwen	0.1802	0.1063	0.1722	0.1824
Аудиопризнаки + delta_ppl, ratio_ppl	Qwen	0.1645	0.0801	0.1392	0.1440
Аудиопризнаки + ppl гипотеза	GigaChat3	0.1625	0.1076	0.1493	0.1872
Аудиопризнаки + delta_ppl, ratio_ppl	GigaChat3	0.1390	0.0843	0.1209	0.1488
Аудиопризнаки + ppl гипотеза	Yandex	0.1527	0.1051	0.1419	0.1890
Аудиопризнаки + delta_ppl, ratio_ppl	Yandex	**0.1274**	0.0860	**0.1146**	0.1512
Аудиопризнаки + ppl гипотеза	Ruadapt	0.1639	0.1072	0.1592	0.1923
Аудиопризнаки + delta_ppl, ratio_ppl	Ruadapt	0.1449	0.0843	0.1289	0.1517

Таким образом, можно сделать следующие выводы.

- Использование любых признаков с перплексией дает статистически значимый прирост точности предсказания качества распознавания речи по сравнению с аудиометриками.
- Признаки с перплексией дают улучшения результатов в 1.5–2 раза по сравнению с базовой моделью на основе аудиопризнаков.
- Для моделей Whisper и Conformer получаются различные результаты по используемым языковым моделям, что связано с разными типами ошибок моделей, а именно: модель Whisper более склона к галлюцинациям, чем Conformer, при этом она сохраняет связность текста, более устойчива к фонетическим ошибкам спикеров.

Полученные результаты подчеркивают, что выбор языковой модели для предсказания качества распознавания речи должен основываться на типе ожидаемых ошибок.

3.1. Отбор признаков

Для выявления наиболее значимых признаков была выполнена процедура отбора признаков методом SHAP со стандартным вычислением значений из фреймворка CatBoost для каждой пары модель – датасет отдельно (см. табл. 3) и подбор признаков в среднем лучший для всех пар (см. табл. 4).

Табл. 3. Отбор признаков по всем возможным комбинациям для каждого набора по метрике RMSE.

Эксперимент	Fleurs Conformer	Fleurs Whisper	LS Conformer	LS Whisper
Аудиопризнаки	0.1976	0.1078	0.1891	0.2163
Аудиопризнаки + delta_ppl, ratio_ppl (все LLM)	**0.1009**	0.0759	**0.0933**	**0.1040**
10 отобранных признаков	0.10397	**0.0649**	0.0949	0.1041
5 отобранных признаков	0.1112	0.0666	0.1025	0.1137
3 отобранных признака	0.1169	0.0683	0.1127	0.1203
1 отобранный признак	0.1449	0.0711	0.1587	0.1607

Табл. 4. Отбор признаков общих для всех наборов по метрике RMSE.

Эксперимент	Выбранные признаки	Fleurs Conformer	Fleurs Whisper	LS Conformer	LS Whisper
Аудиопризнаки		0.1976 ± 0.0049	0.1078 ± 0.0073	0.1891 ± 0.0014	0.2163 ± 0.0013
10 отобранных	nisqa_loud, wada_snr, delta_ppl_gpt2, ratio_ppl_gpt2, ratio_ppl_qwen, яdelta_ppl_gigachat3, ratio_ppl_gigachat3, delta_ppl_yandex, ratio_ppl_yandex, ratio_ppl_ruadapt	**0.1163**	0.0812	**0.0978**	**0.1084**
5 отобранных	delta_ppl_gpt2, ratio_ppl_gpt2, ratio_ppl_qwen, ratio_ppl_gigachat3, delta_ppl_yandex	0.1304	0.1875	0.1095	0.1178

3 отобранных	delta_ppl_gpt2, ratio_ppl_qwen, delta_ppl_yandex	0.1417	0.2261	0.1213	0.1270
1 отобранный	ratio_ppl_gigachat3	0.1908	0.3821	0.1645	0.2079
8 выбранных вручную	nisqa_loud, wada_snr, delta_ppl_gpt2, ratio_ppl_gpt2, delta_ppl_gigachat3, ratio_ppl_gigachat3, delta_ppl_yandex, ratio_ppl_yandex	0.1186	0.0808	0.1024	0.1122
6 выбранных вручную	nisqa_loud, wada_snr, delta_ppl_gpt2, ratio_ppl_gpt2, delta_ppl_yandex, ratio_ppl_yandex	0.12001	**0.0766**	0.1084	0.1156
4 выбранных вручную	nisqa_loud, wada_snr, ratio_ppl_gpt2, ratio_ppl_yandex	0.1254	0.0772	0.1135	0.1199

На основе результатов, представленных в табл. 3 и 4, можно сделать следующие выводы:

- добавление признаков, основанных на отношениях перплексий языковых моделей (delta_ppl, ratio_ppl), к аудиопризнакам приводит к значительному улучшению результатов (снижению значения метрики) во всех экспериментах;
- среди всех LLM стабильно высокую важность демонстрируют признаки, полученные с помощью моделей GPT-2 и Yandex GPT (например, ratio_ppl_gpt2, delta_ppl_gpt2, ratio_ppl_yandex, delta_ppl_yandex). Они занимают верхние строчки практически во всех экспериментах.

3.2. Переносимость модели

Для построенных моделей прогнозирования качества распознавания речи важным свойством является переносимость между различными датасетами. Под переносимостью понимается обучение модели на одном наборе данных с последующей оценкой качества на другом. Это позволяет применять одну и ту же модель к различным корпусам без отдельной настройки под каждый из них. В связи с этим были проведены эксперименты по переносимости (см. табл. 5).

Табл. 5. Переносимость между наборами данных, отбор признаков по метрике RMSE.

Эксперимент	Train Fleurs Conformer – Test LS Conformer	Train LS Conformer - Test Fleurs Conformer	Train Fleurs Whisper – Test LS Whisper	Train LS Whisper – Test Fleurs Whisper
Аудиопризнаки	0.2287	0.2386	0.4187	0.1932
Аудиопризнаки + delta_ppl, ratio_ppl (все LLM)	**0.1049**	**0.1178**	0.2279	0.0580
10 отобранных признаков	0.1061	0.1184	0.2228	**0.0556**
5 отобранных признаков	0.1168	0.1210	**0.2196**	0.0630
3 отобранных признаков	0.1250	0.1421	0.2323	0.0634

В эксперименте "Train Fleurs -> Test LS" для модели Conformer ошибка (WER) снизилась более чем в два раза — с 0.2287 до 0.1049. Лингвистические признаки эффективно компенсируют различие данных между доменами, делая модель более универсальной.

Положительный эффект от добавления признаков LLM наблюдается для обеих архитектур (Conformer и Whisper). Наибольшее относительное улучшение заметно для Conformer, однако для Whisper использование признаков LLM также позволяет достичь минимальной ошибки (0.0556 в эксперименте "Train LS -> Test Fleurs" для 10 отобранных признаков).

ЗАКЛЮЧЕНИЕ

В рамках проведенного исследования исследована задача по прогнозированию WER современных систем автоматического распознавания речи (ASR) на наборе данных с акустическими деградациями, характерными для нестабильного интернет-соединения:

• для количественной оценки степени искажений был апробирован метод прогнозирования WER, основанный на анализе совокупности акустических характеристик сигнала и автоматических нейросетевых метрик;

• для метода прогнозирования WER были исследованы признаки перплексии для современных языковых моделей;

• добавление признаков ratio и delta для языковых моделей позволило обеспечить переносимость качества прогноза при кросс доменном обучении модели прогнозирования WER.

Благодарности

Исследование выполнено в рамках Государственного задания МГУ имени М. В. Ломоносова.

СПИСОК ЛИТЕРАТУРЫ

1. *Wang Q., Sun J., Peng Y., Watanabe S.* Improving Multilingual Speech Models on ML-SUPERB 2.0: Fine-tuning with Data Augmentation and LID-Aware CTC // Proc. Interspeech. 2025. P. 2088–2092.
https://doi.org/10.21437/Interspeech.2025-2377

2. *Nishida T. et al.* Description and Discussion on DCASE 2025 Challenge Task 2: First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring // arXiv preprint arXiv: 2506.10097. 2025.
URL: https://arxiv.org/abs/2506.10097

3. *Valladares-Poncela A., Fraga-Lamas P., Fernández-Caramés T.M.* On-Device Automatic Speech Recognition for IIoT and Extended Reality Industrial Metaverse Applications // Engineering Proceedings. 2024. Vol. 82. Article 3.
https://doi.org/10.3390/ecsa-11-20466

4. *Полевой А.В., Корухова Ю.С.* Об одном подходе к формальной верификации нейросетевых моделей // Научный сервис в сети Интернет: труды XXVI Всероссийской научной конференции (23–25 сентября 2024 г., онлайн). М.: ИПМ им. М.В.Келдыша, 2024. С. 223–235. https://doi.org/10.20948/abrau-2024-11

5. *Jakhar N., Srivastava S., Baby A.* A Unified Approach to Multilingual Automatic Speech Recognition with Improved Language Identification for Indic Languages // Proc. Interspeech. 2024. P. 3949–3953.

6. *Shah M.A., Noguero D.S., Heikkila M.A., Raj B., Kourtellis N.* Speech robust bench: A robustness benchmark for speech recognition // arXiv preprint. arXiv: 2403.07937. 2024. URL: https://arxiv.org/abs/2403.07937.

7. *Baevski A., Zhou Y., Mohamed A., Auli M.* wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 12449–12460.

8. *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.

9. *Tsilfidis A., Mporas I., Mourjopoulos J., Fakotakis N.* Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing // Computer Speech & Language. 2013. Vol. 27, No. 1. P. 380–395.

10. *Loweimi E., Ahadi S.M., Drugman T., Loveymi S.* On the importance of pre-emphasis and window shape in phase-based speech recognition // Advances in Nonlinear Speech Processing: 6th International Conference, NOLISP 2013. 2013. P. 160–167.

11. *Yin S., Liu C., Zhang Z., Lin Y., Wang D., Tejedor J., Zheng T.F., Li Y.* Noisy training for deep neural networks in speech recognition // EURASIP Journal on Audio, Speech, and Music Processing. 2015. Vol. 2015. P. 1–14.

12. *Kim C., Stern R.M.* Power-normalized cepstral coefficients (PNCC) for robust speech recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016. Vol. 24, No. 7. P. 1315–1329.

13. *Li J., Deng L., Gong Y., Haeb-Umbach R.* An overview of noise-robust automatic speech recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014. Vol. 22, No. 4. P. 745–777.

14. *Stern R.M., Morgan N. Hearing is believing:* Biologically inspired methods for robust automatic speech recognition // IEEE Signal Processing Magazine. 2012. Vol. 29, No. 6. P. 34–43.

15. *Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I.* Robust speech recognition via large-scale weak supervision // International Conference on Machine Learning. 2023. P. 28492–28518.

16. *Bouchakour L., Debyeche M.* Noise-robust speech recognition in mobile network based on convolution neural networks // International Journal of Speech Technology. 2022. Vol. 25. P. 1–9.

17. *Duarte J., Colcher S.* Noise-Robust Automatic Speech Recognition: A Case Study for Communication Interference // Journal on Interactive Systems. 2024. Vol. 15. P. 670–681.

18. *Chen G., O'Shaughnessy D., Tolba H.* A performance investigation of noisy voice recognition over IP telephony networks // Proc. Interspeech. 2005. P. 2681–2684.

19. *Ali A., Renals S.* Word error rate estimation for speech recognition: e-WER // Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. Vol. 2. P. 20–24.

20. *Ali A., Renals S.* Word error rate estimation without ASR output: e-WER2 // arXiv preprint arXiv: 2008.03403. 2020.

21. *Chowdhury S.A., Ali A.* Multilingual word error rate estimation: e-WER3 // ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing. 2023. P. 1–5.

22. *Litman D., Hirschberg J., Swerts M.* Predicting automatic speech recognition performance using prosodic cues // 1st Meeting of the North American Chapter of the Association for Computational Linguistics. 2000. P. 1–8.

23. *Gallardo L.F., Möller S., Beerends J.* Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores // Proc. Interspeech 2017. 2017. P. 2939–2943.

24. *Beerends J.G., Schmidhuber C., Berger J., Obermann M., Ullmann R., Pomy J., Kevhl M.* Perceptual Objective Listening Quality Assessment (POLQA), the Third Generation ITU-T Standard for End-to-End Speech Quality Measurement. Part I: Temporal Alignment // AES: Journal of the Audio Engineering Society. 2013. Vol. 61, No. 6. P. 366–384.

25. *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. P. 29–44.

26. *Mittag G., Naderi B., Chehadi A., Möller S.* NISQA: A deep CNN–self-attention model for multidimensional speech quality prediction with crowdsourced datasets // Proc. Interspeech 2021. 2021. P. 1–5.
https://doi.org/10.21437/Interspeech.2021-299

27. *Kim C., Stern R.* Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis // Proc. Interspeech 2008. 2008. P. 2598–2601. https://doi.org/10.21437/Interspeech.2008-644

28. *Subakan C., Ravanelli M., Cornell S., Grondin F.* REAL-M: Towards Speech Separation on Real Mixtures // arXiv preprint arXiv:2110.10812. 2021.

29. *Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I.* Language models are unsupervised multitask learners // OpenAI Blog. 2019.

30. *Yang A. et al.* Qwen2 technical report // arXiv preprint. arXiv: 2407.10671. 2024.

31. ai-sage/GigaChat3-10B-A1.8B // Hugging Face.
URL: https://huggingface.co/ai-sage/GigaChat3-10B-A1.8B

32. *Yandex* YandexGPT-5-Lite-8B-instruct // Hugging Face.
URL: https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct

33. *RefalMachine* RuadaptQwen3-8B-Hybrid // Hugging Face.
URL: https://huggingface.co/RefalMachine/RuadaptQwen3-8B-Hybrid

34. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.* CatBoost: unbiased boosting with categorical features // arXiv preprint arXiv:1706.09516. 2019.

35. *Conneau A. et al.* FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech // arXiv preprint. arXiv: 2205.12446. 2022.

36. *Panayotov V., Chen G., Povey D., Khudanpur S.* Librispeech: An ASR corpus based on public domain audio books // ICASSP 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. 2015. P. 5206–5210.

AUTOMATIC SPEECH RECOGNITION QUALITY PREDICTION BASED ON LARGE LANGUAGE MODELS

A. V. Polevoi[0009-0000-2216-0468]

Lomonosov Moscow State University, Moscow, Russian Federation

polevoianton@bk.ru

Abstract

Despite the rapid development in the field of automatic speech recognition (ASR) systems, the recognition quality remains poor for audio recordings with acoustic degradation (music, crowd shouts, sounds of machinery, etc.). This becomes especially important when implementing models in critical areas that are characterized by increased attention to control and reliability of the results (aviation, medicine, autopilots, etc.). Error in such areas is very expensive and the use of models becomes impossible, even if the error occurs rarely. To reduce risks, it is advisable to evaluate the expected recognition quality in advance.

This article proposes an approach to predicting the Word Error Rate (WER) based on the acoustic characteristics of the signal and calculating the perplexity of language models. The proposed method involves the creation of diverse sets of audio data by applying various types of acoustic observations to pure speech samples at various levels of quality and intelligibility. Unlike previous studies, a complete set of speech features is extracted and analyzed: prediction of the value of the signal-to-noise ratio (SNR), neural network sound quality metrics (NISQA, etc.) as well as the perplexity of the text of the ASR hypothesis using the language model as an additional feature for training a unified model.

Experiments are being conducted using modern speech recognition architectures to demonstrate the effectiveness of the proposed method in predicting WER in various acoustic conditions. It is shown that the inclusion of perplexity significantly improves the quality of the WER prediction, in particular for data where acoustic features are weakly correlated with recognition errors. The results are applicable for automatic evaluation of the expected quality of speech recognition and filtering of audio inputs.

***Keywords**: WER prediction, acoustic degradation in speech recognition, perplexia, ASR confidence.*

REFERENCES

1. *Wang Q., Sun J., Peng Y., Watanabe S.* Improving Multilingual Speech Models on ML-SUPERB 2.0: Fine-tuning with Data Augmentation and LID-Aware CTC // Proc. Interspeech. 2025. P. 2088–2092. https://doi.org/10.21437/Interspeech.2025-2377.

2. *Nishida T. et al.* Description and Discussion on DCASE 2025 Challenge Task 2: First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring // arXiv preprint arXiv:2506.10097. 2025. https://doi.org/10.48550/arXiv.2506.10097

3. *Valladares-Poncela A., Fraga-Lamas P., Fernández-Caramés T.M.* On-Device Automatic Speech Recognition for IIoT and Extended Reality Industrial Metaverse Applications // Engineering Proceedings. 2024. Vol. 82. Article 3. https://doi.org/10.3390/ecsa-11-20466

4. *Polevoj A.V., Koruhova Yu.S.* Ob odnom podhode k formal'noj verifikacii nejrosetevyh modelej // Nauchnyj servis v seti Internet: trudy XXVI Vserossijskoj nauchnoj konferencii (23–25 sentyabrya 2024 g., onlajn). M.: IPM im. M.V. Keldysha, 2024. S. 223–235. https://doi.org/10.20948/abrau-2024-11

5. *Jakhar N., Srivastava S., Baby A.* A Unified Approach to Multilingual Automatic Speech Recognition with Improved Language Identification for Indic Languages // Proc. Interspeech. 2024. C. 3949–3953.

6. *Shah M.A., Noguero D.S., Heikkila M.A., Raj B., Kourtellis N.* Speech robust bench: A robustness benchmark for speech recognition // arXiv preprint arXiv:2403.07937. 2024. https://doi.org/10.48550/arXiv.2403.07937

7. *Baevski A., Zhou Y., Mohamed A., Auli M.* wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 12449–12460.

8. *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.

9. *Tsilfidis A., Mporas I., Mourjopoulos J., Fakotakis N.* Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing // Computer Speech & Language. 2013. Vol. 27. No. 1. P. 380–395.

10. *Loweimi E., Ahadi S.M., Drugman T., Loveymi S.* On the importance of pre-emphasis and window shape in phase-based speech recognition // Advances in Nonlinear Speech Processing: 6th International Conference, NOLISP 2013. 2013. P. 160–167.

11. *Yin S., Liu C., Zhang Z., Lin Y., Wang D., Tejedor J., Zheng T.F., Li Y.* Noisy training for deep neural networks in speech recognition // EURASIP Journal on Audio, Speech, and Music Processing. 2015. Vol. 2015. P. 1–14.

12. *Kim C., Stern R.M.* Power-normalized cepstral coefficients (PNCC) for robust speech recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016. Vol. 24. No. 7. P. 1315–1329.

13. *Li J., Deng L., Gong Y., Haeb-Umbach R.* An overview of noise-robust automatic speech recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014. Vol. 22. No. 4. P. 745–777.

14. *Stern R.M., Morgan N. Hearing is believing:* Biologically inspired methods for robust automatic speech recognition // IEEE Signal Processing Magazine. 2012. Vol. 29. No. 6. P. 34–43.

15. *Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I.* Robust speech recognition via large-scale weak supervision // International Conference on Machine Learning. 2023. P. 28492–28518.

16. Bouchakour L., Debyeche M. Noise-robust speech recognition in mobile network based on convolution neural networks // International Journal of Speech Technology. 2022. Vol. 25. P. 1–9.

17. *Duarte J., Colcher S.* Noise-Robust Automatic Speech Recognition: A Case Study for Communication Interference // Journal on Interactive Systems. 2024. Vol. 15. P. 670–681.

18. *Chen G., O'Shaughnessy D., Tolba H.* A performance investigation of noisy voice recognition over IP telephony networks // Proc. Interspeech. 2005. P. 2681–2684.

19. *Ali A., Renals S.* Word error rate estimation for speech recognition: e-WER // Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. Vol. 2. P. 20–24. https://doi.org/10.18653/v1/P18-2004

20. *Ali A., Renals S.* Word error rate estimation without ASR output: e-WER2 // arXiv preprint arXiv:2008.03403. 2020.

21. *Chowdhury S.A., Ali A.* Multilingual word error rate estimation: e-WER3 // ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing. 2023. P. 1–5.

22. *Litman D., Hirschberg J., Swerts M.* Predicting automatic speech recognition performance using prosodic cues // 1st Meeting of the North American Chapter of the Association for Computational Linguistics. 2000. P. 1–8.

23. *Gallardo L.F., Möller S., Beerends J.* Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores // Proc. Interspeech 2017. 2017. P. 2939–2943.

24. Beerends J.G., Schmidhuber C., Berger J., Obermann M., Ullmann R., Pomy J., Kevhl M. Perceptual Objective Listening Quality Assessment (POLQA), the Third Generation ITU-T Standard for End-to-End Speech Quality Measurement. Part I: Temporal Alignment // AES: Journal of the Audio Engineering Society. 2013. Vol. 61. No. 6. P. 366–384.

25. *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. P. 29–44.

26. *Mittag G., Naderi B., Chehadi A., Möller S.* NISQA: A deep CNN–self-attention model for multidimensional speech quality prediction with crowdsourced datasets // Proc. Interspeech 2021. 2021. P. 1–5.
https://doi.org/10.21437/Interspeech.2021-299

27. *Kim C., Stern R.* Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis // Proc. Interspeech 2008. 2008. P. 2598–2601. https://doi.org/10.21437/Interspeech.2008-644

28. *Subakan C., Ravanelli M., Cornell S., Grondin F.* REAL-M: Towards Speech Separation on Real Mixtures // arXiv preprint arXiv:2110.10812. 2021.

29. *Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I.* Language models are unsupervised multitask learners // OpenAI Blog. 2019.

30. *Yang A. et al.* Qwen2 technical report // arXiv preprint arXiv:2407.10671. 2024.

31. *ai-sage* GigaChat3-10B-A1.8B // Hugging Face. URL: https://huggingface.co/ai-sage/GigaChat3-10B-A1.8B

32. *Yandex* YandexGPT-5-Lite-8B-instruct // Hugging Face. URL: https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct

33. *RefalMachine* RuadaptQwen3-8B-Hybrid // Hugging Face. URL: https://huggingface.co/RefalMachine/RuadaptQwen3-8B-Hybrid

34. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.* CatBoost: unbiased boosting with categorical features // arXiv preprint. 2019. https://doi.org/10.48550/arXiv.1706.09516

35. *Conneau A. et al.* FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech // arXiv preprint arXiv:2205.12446. 2022.

36. *Panayotov V., Chen G., Povey D., Khudanpur S.* Librispeech: An ASR corpus based on public domain audio books // ICASSP 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. 2015. P. 5206–5210.

СВЕДЕНИЯ ОБ АВТОРЕ

ПОЛЕВОЙ Антон Вячеславович – аспирант; Московский государственный университет имени М. В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Российская Федерация.

POLEVOI Anton Vyacheslavovich – postgraduate student at Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Moscow, Russian Federation.

email: polevoianton@bk.ru

ORCID: 0009-0000-2216-0468

Материал поступил в редакцию 5 апреля 2026 года