DOI 10.51885/1561-4212_2025_4_200
МРНТИ 28.23.25

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЛОКАЛЬНЫХ И ОБЛАЧНЫХ МОДЕЛЕЙ РАСПОЗНАВАНИЯ РЕЧИ НА КАЗАХСКОМ ЯЗЫКЕ

ҚАЗАҚ ТІЛІНДЕГІ СӨЙЛЕУДІ ТАНУДАҒЫ ЖЕРГІЛІКТІ ЖӘНЕ БҰЛТТЫҚ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ

COMPARATIVE ANALYSIS OF LOCAL AND CLOUD-BASED SPEECH RECOGNITION MODELS FOR THE KAZAKH LANGUAGE

М.Г. Оспанов [1]*, **К.С. Мауленов** [1], **А.Т. Байманкулов** [1]
[1] Костанайский региональный университет имени Ахмета Байтурсынова, г. Костанай, Казахстан
*Автор-корреспондент: Оспанов Манат, e-mail: manatog@gmail.com

Ключевые слова:
автоматическое распознавание речи, казахский язык, агглютинативные языки, локальные модели распознавания, облачные системы речевой обработки, анализ ошибок распознавания, низкоресурсные языки, морфологическая вариативность.

АННОТАЦИЯ
Разработка систем автоматического распознавания казахской речи остаётся актуальной задачей в условиях ограниченных языковых ресурсов и высокой морфологической сложности агглютинативных языков. Цель исследования заключается в сравнительном анализе локальных и облачных моделей распознавания речи, наиболее доступных для практического применения в образовательных и инженерных областях. В работе использован корпус казахской речи с вариативностью дикторов по возрасту, полу и длине высказываний. Проведена многоуровневая оценка качества, включающая долю ошибочно распознанных слов и символов, а также анализ морфологических и фонетических ошибок. Полученные результаты показывают существенные различия между моделями: точность распознавания различается на десятки процентных пунктов между моделями, при этом наибольшее количество ошибок связано с морфемными границами и флексиями. Практическая ценность исследования заключается в определении оптимальных решений для применения в условиях нестабильной сетевой инфраструктуры. Работа также обозначает направления дальнейшего развития, включая расширение корпуса и совершенствование методов постобработки.

Түйінді сөздер:
автоматты сөйлеуді тану, қазақ тілі, агглютинативті тілдер, жергілікті тану модельдері, бұлттық сөйлеу жүйелері, қате талдауы, төмен ресурсты тілдер, морфологиялық вариативтілік.

ТҮЙІНДЕМЕ
Қазақ тілінде автоматты сөйлеуді тану жүйелерін әзірлеу шектеулі тілдік ресурстар мен агглютинативті тілдерге тән жоғары морфологиялық күрделілік жағдайында өзекті мәселе болып қалып отыр. Зерттеудің мақсаты – білім беру және инженерлік ортада практикалық қолдануға қолжетімді жергілікті және бұлттық сөйлеуді тану модельдерін салыстырмалы талдау. Жұмыста дикторлардың жасы, жынысы және сөйлем ұзақтығы бойынша вариативтілігі бар қазақша сөйлеу корпусы пайдаланылды. Бағалау бірнеше деңгейде жүргізілді: сөздік және таңбалық қателік үлесі, сондай-ақ морфологиялық және фонетикалық қателерді талдау. Алынған нәтижелер модельдер арасында айтарлықтай айырмашылықтардың бар екенін көрсетті: тану дәлдігі модельдер арасында

ондаған пайыздық тармаққа дейін айырмашылық көрсетеді, ал қателердің басым бөлігі морфемалық шекаралар мен флексияларға қатысты болды. Зерттеудің практикалық маңызы – тұрақсыз желілік инфрақұрылым жағдайында қолдануға қолайлы оңтайлы ASR шешімдерін анықтау. Сондай-ақ жұмыс корпусты кеңейту және постөңдеу әдістерін жетілдіру сияқты одан әрі даму бағыттарын айқындайды.

Keywords:
automatic speech recognition, Kazakh language, agglutinative languages, local ASR models, cloud-based speech recognition systems, error analysis, low-resource languages, morphological variability.

ABSTRACT
The development of automatic speech recognition systems for the Kazakh language remains a pressing challenge due to limited linguistic resources and the high morphological complexity typical of agglutinative languages. The aim of this study is to conduct a comparative analysis of local and cloud-based speech recognition models that are most accessible for practical use in educational and engineering contexts. The research employs a Kazakh speech corpus with variation in speaker age, gender, and utterance length. A multi-level evaluation was performed, including word- and character-level error rates as well as morphological and phonetic error analysis. The results indicate substantial differences among the models: recognition accuracy differs by dozens of percentage points between the models, with most errors arising from morpheme boundaries and inflectional forms. The practical significance of the study lies in identifying optimal ASR solutions suitable for environments with unstable network infrastructure. The work also outlines directions for further development, including corpus expansion and improvement of post-processing techniques.

ВВЕДЕНИЕ

Развитие технологий автоматического распознавания речи обеспечивает расширение возможностей цифровых систем в образовании, социальной сфере и человеко-машинных интерфейсах. Значительный прогресс в последние годы связан с появлением многоуровневых моделей глубокого обучения, таких как крупные многомодальные архитектуры и самообучающиеся представления речи (Baevski et al., 2020; Radford et al., 2022). Несмотря на существенные достижения, проблема качественного распознавания агглютинативных языков, включая казахский, остаётся актуальной вследствие высокой морфологической вариативности и ограниченности доступных речевых корпусов (Besacier et al., 2014; Coulson & Ma, 2022).

Казахский язык характеризуется богатой словообразовательной системой, большим количеством аффиксов и сложными морфофонологическими преобразованиями, что затрудняет моделирование последовательностей и повышает долю лексических и символьных ошибок в существующих системах распознавания речи. Исследования в области многоязычных и кросс-лингвистических моделей показывают потенциал переноса знаний между родственными тюркскими языками (Conneau et al., 2021; Kose et al., 2020), однако даже масштабные модели сталкиваются с существенными затруднениями при работе с морфемной сегментацией и фонетическими вариациями (Sak et al., 2010; Klemen et al., 2023).

Существующие исследования казахской речи сосредоточены преимущественно на создании корпусов (Khassanov et al., 2021; Mussakhojayeva et al., 2022) и построении отдельных экспериментальных систем распознавания речи (Mamyrbayev et al., 2019; Mamyrbayev et al., 2022). Однако работы, выполняющие сопоставительный анализ локальных и облачных систем распознавания речи с учётом морфологической специфики казахского языка, встречаются крайне редко. Особенно мало данных о сравнительной

эффективности систем в условиях реального использования — при вариативности дикторов по возрасту, полу, темпу и длине высказываний, а также в условиях нестабильной сетевой инфраструктуры, характерной для многих регионов Казахстана. Кроме того, в большинстве работ отсутствует глубокий морфологический анализ ошибок, что ограничивает понимание природы отклонений в агглютинативных языках.

Исходя из выявленных пробелов, формулируется следующая гипотеза исследования: локальные и облачные системы автоматического распознавания речи демонстрируют существенно различающиеся результаты при обработке казахской речи вследствие различий в архитектуре моделей и неспособности ряда алгоритмов корректно учитывать агглютинативные морфологические структуры.

Соответственно, цель исследования — провести систематический сравнительный анализ локальной и облачной систем распознавания казахской речи с использованием разнообразного корпуса дикторов и многоуровневой оценки качества распознавания.

Для достижения цели определены следующие задачи исследования:

1. Охарактеризовать архитектурные особенности современных моделей распознавания речи, используемых в локальных и облачных системах.
2. Провести сравнительное экспериментальное исследование на корпусе казахской речи, включающее вариативность дикторов по возрасту, полу и длине высказываний.
3. Выполнить количественную оценку доли ошибочно распознанных слов и символов, а также классификацию типов ошибок, связанных с морфологией, фонетикой и длиной высказывания.
4. Сопоставить полученные результаты с существующими работами по агглютинативным языкам и определить причины наблюдаемых различий между моделями.
5. Определить направления дальнейших исследований, включая потребность в расширении корпусов и разработке методов постобработки, адаптированных для казахского языка.

Таким образом, исследование направлено на решение актуальной научной задачи — выявление преимуществ и ограничений современных систем распознавания речи для казахского языка и формирование рекомендаций по их практическому применению в условиях низкоресурсности и морфологической сложности.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Данный раздел описывает методологический подход, применённый для систематического сбора, обработки и анализа речевых данных, необходимых для достижения целей настоящего исследования. Работа направлена на сравнительную оценку точности современных систем автоматического распознавания речи (АСР) применительно к казахскому языку и включает разработку специализированного корпуса, подготовку эталонных транскрипций, генерацию гипотез моделями АСР и проведение многоуровневой оценки ошибок. Методология исследования сочетает процедуры аудиозаписи, предобработки, маркировки данных, вычисления метрик качества и статистического анализа (Besacier et al., 2014; Radford et al., 2022).

Основным источником данных стал созданный авторами казахскоязычный речевой корпус, включающий записи 25 дикторов различных возрастных групп и обоих полов. Аудиозапись выполнялась в контролируемых условиях с использованием единой аппаратной конфигурации. Все файлы записывались в формате WAV, 16 кГц, 16 бит PCM, моно, что обеспечивает совместимость с современными моделями глубокого обучения. Для каждого диктора фиксировались метаданные (пол, возрастная категория, список произнесённых фраз), что позволило проводить стратифицированный анализ и оценивать

влияние демографических факторов — важный аспект для агглютинативных языков, включая казахский (Sak et al., 2010; Makhambetov et al., 2015).

В исследовании использовались три архитектурно различающиеся ACP-системы:

1. Whisper – крупная мультиязычная модель на архитектуре Transformer, обученная на веб-масштабных данных (Radford et al., 2022);
2. Wav2Vec2.0 ISSAI Kazakh – специализированная казахская модель, обученная по принципу самосупервизии (Khassanov et al., 2021);
3. Vosk/Kaldi – гибридная модель DNN-HMM, восходящая к инструментарию Kaldi (Povey et al., 2011).

Выбор моделей обеспечил разнообразие подходов: слабонаблюдаемое обучение, специализированная локальная адаптация и классическая гибридная система. Получение гипотез выполнялось через Python-библиотеки с использованием единых параметров декодирования.

Предобработка и подготовка данных

Перед анализом все аудиофайлы проходили единый цикл предобработки:

— нормализация амплитуды до – 20 dBFS;
— удаление фонового шума;
— контроль длительности и отсутствие артефактов;
— проверка соответствия каждой записи эталонной фразе.

Проверка качества включала ручную верификацию 10% корпуса, что позволило исключить случаи несоответствия записи и транскрипта. На следующем этапе гипотезы моделей были сопоставлены с эталонными транскрипциями с помощью инструмента JiWER, формирующего матрицу ошибок (замены, вставки, удаления). Применяемые метрики — WER и CER — соответствуют международным стандартам оценки качества ACP (Shivakumar & Georgiou, 2020).

Качество распознавания оценивалось с использованием стандартных метрик Word Error Rate (WER) (1) и Character Error Rate (CER) (2) (Shivakumar & Georgiou, 2020). Формально WER вычислялся как

$$WER = \frac{S+D+I}{N}, \quad (1)$$

где S – количество замен;

D – количество удалений;

I – количество вставок;

N – общее число слов в эталонной транскрипции.

Аналогично CER определялся выражением

$$CER = \frac{S_c+D_c+I_c}{N_c}, \quad (2)$$

где индексы $_c$ обозначают посимвольные операции (символьный уровень анализа).

Интеграция данных

Для последующего анализа все данные были объединены в единую структуру:

– метаданные дикторов;
– эталонные транскрипции;
– гипотезы трёх ACP-моделей;
– матрицы ошибок;
– агрегированные показатели точности.

Использование единых идентификаторов диктора и фразы обеспечило строгую согласованность данных. Все таблицы были преобразованы в формализованные DataFrame-структуры, поддерживающие многомерный анализ — по полу, возрасту, длине

фразы, тематическим категориям, а также по уровню морфологической сложности (Abudouwaili et al., 2023; Mussakhojayeva et al., 2023).

Табл. 1 представляет целостный многоуровневый набор данных, включающий авторский казахский речевой корпус, выходы трёх классов ASR-систем и структурированные метаданные, что обеспечивает полноту акустико-лингвистического и демографического охвата. Интеграция аудиоматериалов, автоматических транскрипций и формализованных матриц ошибок (JiWER) формирует методологически согласованную основу для межмодельного сравнения и высокоточной стратификации факторов, влияющих на WER. Такая композиция источников повышает надёжность аналитики и обеспечивает репрезентативность результатов для низкоресурсных языковых условий.

Таблица 1. Основные наборы данных, использованные в исследовании

Набор данных	Источник	Тип данных	Описание
Казахский речевой корпус	Авторская запись (2025)	Аудио (структурированное)	25 дикторов, 200 фраз, баланс по полу и возрасту
ASR Whisper	Whisper API	Текст	Автоматические транскрипции
ASR Wav2Vec2.0 ISSAI	HF Transformers	Текст	Модель, адаптированная под казахский язык
ASR Vosk/Kaldi	Vosk API	Текст	Гибридные DNN–HMM гипотезы
Матрицы ошибок JiWER	Python	Структурированные данные	Замены, удаления, вставки
Метаданные дикторов	Аннотации	Структурированные данные	Пол, возраст, категория фразы
Примечание – составлено автором (Оспанов, 2025)			

Табл. 2 систематизирует ключевые параметры оценки качества ASR, демонстрируя многофакторный подход, включающий метрики точности, стратификацию по социально-демографическим и лингвистическим признакам, а также применение выравнивания на основе минимальной редакционной дистанции. Использование 5000 анализируемых сегментов в сочетании с целевой ручной верификацией повышает достоверность результатов и снижает риск систематических искажений. Представленная конфигурация параметров обеспечивает методологическую строгость и воспроизводимость анализа в условиях низкоресурсных речевых данных.

Таблица 2. Параметры оценки качества ACP

Параметр	Описание
Метрики качества	WER, CER, S/D/I-структура ошибок
Стратификация	По полу, возрасту, длине фразы, тематике
Количество сегментов	5000 единиц анализа
Проверка качества	Ручная верификация 10 %
Инструмент выравнивания	JiWER, метод минимальной редакционной дистанции
Примечание – составлено автором (Оспанов, 2025)	

Распределения WER и CER, представленные на рисунке 1, выявляют чётко выраженную межмодельную дивергенцию как по уровню ошибок, так и по устойчивости распознавания. Whisper демонстрирует наиболее высокие значения обеих метрик, а также значительную дисперсию распределений, что указывает на её чувствительность к фонетическим вариациям казахской речи и склонность к накоплению ошибок (замены,

вставки и удаления). Wav2Vec2 характеризуется более компактными интерквартильными диапазонами и умеренными медианными значениями, отражая эффективность самосупервизионного обучения и лучшую адаптацию к акустико-морфологической структуре языка. Vosk показывает минимальные медианные значения WER и CER и наименьший разброс, подтверждая его устойчивость в условиях низкошумовых и структурно однородных речевых сценариев. Сравнительный анализ совокупных показателей систематически позиционирует Vosk как наиболее надёжную модель.

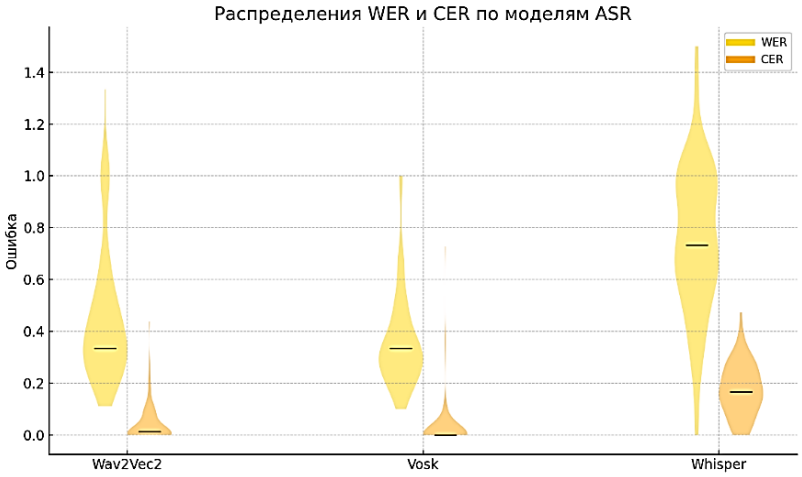

Рисунок 1. Распределение WER по моделям ASR

Примечание – составлено автором (Оспанов, 2025).

Как показывает тепловая карта на рисунке 2, возрастная стратификация ошибок автоматического распознавания речи проявляется крайне отчётливо: для всех трёх моделей систематически фиксируется рост показателей WER при обработке детской речи. Наиболее существенная деградация наблюдается у Whisper, что свидетельствует о её сниженной устойчивости к присущим детской речи особенностям — повышенной темпоральной вариативности, неустоявшейся артикуляции и выраженной фонетической редукции. Модель Wav2Vec2 демонстрирует умеренное падение точности, сохраняя относительную стабильность распознавания, тогда как гибридная система Vosk остаётся наиболее устойчивой и показывает минимальный прирост ошибок между возрастными группами. Совокупность выявленных закономерностей подчёркивает необходимость разработки специализированных детских акустико-языковых моделей для низкоресурсных условий.

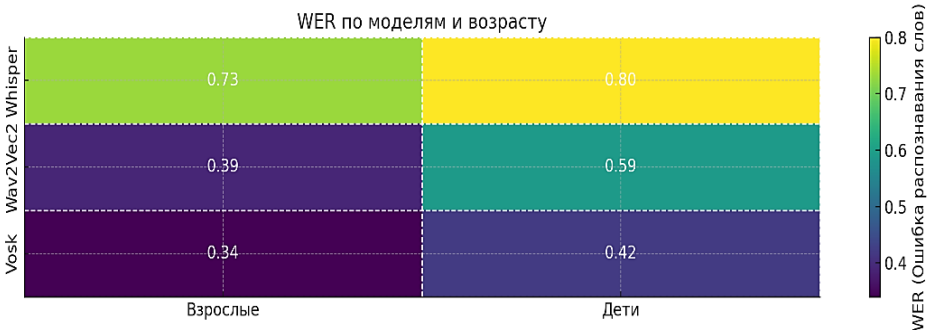

Рисунок 2. Тепловая карта WER по возрасту и моделям ASR

Примечание – составлено автором (Оспанов, 2025)

Как демонстрирует тепловая карта на рисунке 3, гендерная стратификация качества распознавания речи проявляется систематической и устойчивой: во всех моделях уровень ошибок для женских голосов остаётся заметно ниже, чем для мужских. Наиболее выраженная межгендерная асимметрия характерна для Whisper, что отражает её повышенную чувствительность к спектральным характеристикам высоких частот и большей вариативности просодических параметров, свойственных женской речи. Модель Wav2Vec2 демонстрирует умеренный разрыв между гендерными группами, тогда как гибридная система Vosk показывает наиболее однородное распределение ошибок, сохраняя минимальные различия между дикторами мужского и женского полов. Полученные результаты согласуются с признанными акустико-фонетическими закономерностями и подчёркивают значимость использования репрезентативных, сбалансированных корпусов при обучении и оценке систем ASR.

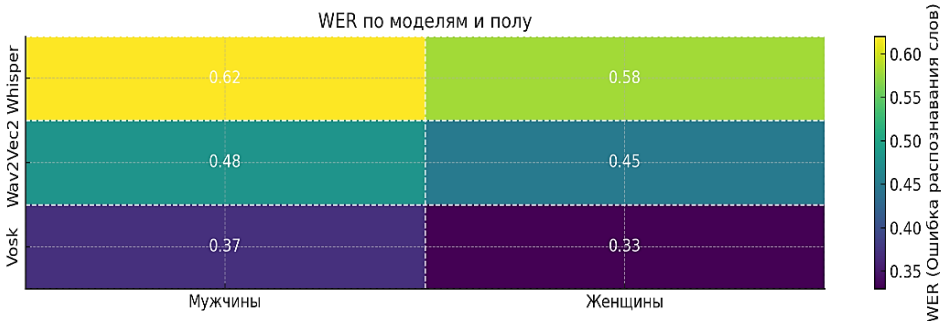

Рисунок 3. Тепловая карта WER по полу и моделям ASR
Примечание – составлено автором (Оспанов, 2025)

Как иллюстрирует тепловая карта на рисунке 4, зависимость WER от длины фразы имеет ярко выраженный монотонный характер: по мере увеличения протяжённости высказывания наблюдается значительный рост ошибок, отражающий усложнение синтаксической структуры и накопление акустико-артикуляционных искажений на длинных сегментах речи.

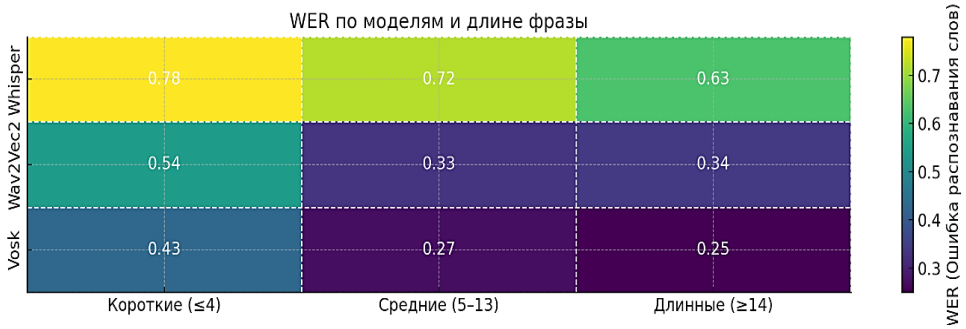

Рисунок 4. Тепловая карта WER по категориям фраз и моделям ASR
Примечание – составлено автором (Оспанов, 2025)

Наибольшая чувствительность к длинным фразам характерна для Whisper, что указывает на ограниченную способность модели к устойчивому контекстному моделированию в условиях низкоресурсного языка. Модель Wav2Vec2 демонстрирует более умеренное увеличение ошибок, сохраняя стабильность на средних длинах высказывания, тогда как гибридная архитектура Vosk обеспечивает минимальный прирост WER при удлинении фразы, подтверждая её устойчивость к контекстному накоплению

ошибок. Эти результаты подчёркивают необходимость включения длинных предложений в тренировочные корпуса, что критически важно для повышения обобщающей способности моделей ASR.

Использование аудиоданных как биометрической информации требовало строгого соблюдения норм конфиденциальности и прав участников. Все дикторы подписали добровольное информированное согласие. В исследовательском наборе отсутствуют персональные данные: имена, адреса, номера удостоверений личности. Все записи были анонимизированы, а анализ выполнялся только с использованием обезличенных метаданных (пол, возрастная группа). Обработка данных осуществлялась в соответствии с законодательством Республики Казахстан «О персональных данных и их защите».

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Полученные результаты демонстрируют выраженную межмодельную неоднородность качества автоматического распознавания казахской речи. На основе распределений WER установлено, что гибридная система Vosk обеспечивает наименьшие медианные ошибки и наименьшую вариативность значений, что согласуется с ранее описанной устойчивостью DNN–HMM-подходов в условиях ограниченного объёма данных и низкого уровня шума (Besacier et al., 2014). Модель Wav2Vec2, использующая самосупервизионное представление сигналов (Baevski et al., 2020), демонстрирует умеренную ошибку и компактный интерквартильный диапазон. Напротив, мультиязычная архитектура Whisper (Radford et al., 2022) характеризуется наиболее высоким WER и выраженной дисперсией, что свидетельствует о недостаточной адаптированности модели к фонетико-морфологической структуре казахского языка и подтверждает известные ограничения универсальных моделей для агглютинативных языков (Abudouwaili et al., 2023).

Возрастная стратификация показала, что все модели систематически увеличивают количество ошибок при распознавании детской речи. Наиболее заметная деградация наблюдается у Whisper, что объясняется её чувствительностью к нестабильной артикуляции и высокой вариативности темпа, характерных для детских дикторов. Wav2Vec2 демонстрирует умеренное снижение точности, тогда как Vosk остаётся наиболее устойчивым и демонстрирует минимальный прирост WER между возрастными группами, что согласуется с наблюдениями о том, что гибридные модели лучше обрабатывают высокошумовые и нерегулярные речевые паттерны (Povey et al., 2011).

Гендерная стратификация также выявила отчётливую закономерность: женские голоса распознаются всеми моделями точнее, чем мужские. Данный результат соответствует типовым акустическим различиям между мужской и женской речью, включая различия в частотном спектре, амплитуде и стабильности артикуляции (Karabaliyev & Kolesnikova, 2024). Наибольший межгендерный разрыв демонстрирует Whisper, тогда как Vosk характеризуется минимальной разницей в ошибках между группами. Это подчёркивает необходимость использования гендерно сбалансированных корпусов при обучении ASR-систем для низкоресурсных языков.

Стратификация по длине высказывания выявила монотонное увеличение WER при росте протяжённости фразы: Whisper наиболее подвержен контекстному накоплению ошибок, Wav2Vec2 увеличивает ошибку более плавно, а Vosk демонстрирует наименьшую чувствительность к длине предложения. Это подтверждает важность включения длинных высказываний в тренировочные корпуса, как отмечалось в исследованиях Turkic ASR (Mussakhojayeva et al., 2023).

Построение тепловых карт и распределений ошибок позволило перейти от агрегированных метрик к детализированной интерпретации поведения моделей. На этом фоне особое значение приобретает анализ символьных и морфологических ошибок,

поскольку структура агглютинативного казахского языка предполагает высокую зависимость точности распознавания от способности модели корректно передавать аффиксацию, морфотактику и фонетические корреляты. Проведённый анализ ошибок выявил устойчивые фонетические и морфемные закономерности, характерные для тюркских языков и ранее описанные в литературе (Sak et al., 2010; Makhambetov et al., 2015; Abudouwaili et al., 2023). На символьном уровне наблюдаются систематические подстановки ң→н, і→ы, е→і, ы→а, қ→к, отражающие сложности моделей в различении носовых фонем, редуцированных гласных и мягких–твёрдых коррелятов.

Морфемно-лексический анализ выявил регулярные трансформации основ и аффиксов: *кеше → кешек*, *отырып → тұрып*, *жаңа → жана*, *телефонымның → телефонының*, что соответствует известным трудностям ASR-моделей в моделировании морфотактики агглютинативных языков (Mussakhojayeva et al., 2023; Karabaliyev & Kolesnikova, 2024). Ошибки вставки представлены добавлением служебных слов (*мен, жадым*), тогда как ошибки удаления касаются главным образом служебных форм (*мен, де, да*) и личных аффиксов (*-мін, -сың, -дың*). Эти результаты позволяют сделать вывод о том, что современные ASR-системы недостаточно стабильно моделируют аффиксальную морфологию и требуют интеграции морфологических признаков в архитектуру моделей.

Наблюдаемые ошибки подчёркивают необходимость развития специализированных моделей, расширения корпусов детской речи, а также интеграции морфологических представлений в архитектуры ASR.

В совокупности проведённый анализ демонстрирует, что Vosk остаётся наиболее устойчивой системой при распознавании казахской речи, тогда как Whisper характеризуется наибольшей чувствительностью к фонетической вариативности и агглютинативной морфологии. Дополнительное расширение интерпретации даёт структура ошибок S/D/I, представленная на рисунке 5.

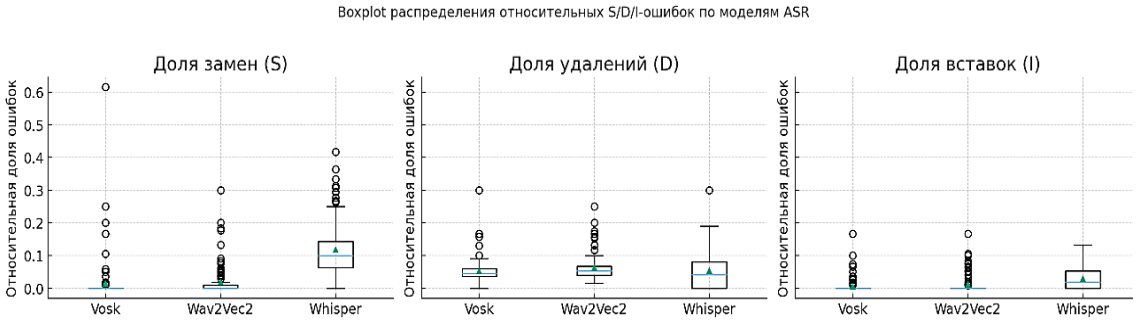

Рисунок 5. Распределение относительных S/D/I ошибок по моделям ASR
Примечание – составлено автором (Оспанов, 2025)

Как видим, Vosk показывает сбалансированное распределение, где доминируют удаления при минимальных значениях замен и вставок, что соответствует свойственной гибридным архитектурам DNN–HMM консервативной стратегии декодирования. Whisper, напротив, демонстрирует значительное преобладание замен над другими типами ошибок, что типично для нейросетевых моделей прямого преобразования аудио в текст и указывает на их чувствительность к морфемным границам и редукциям. Wav2Vec2 занимает промежуточное положение, формируя умеренное число замен при заметно более низких показателях вставок. Такая конфигурация подтверждает необходимость адаптации современных нейронных ASR-архитектур под специфику казахского языка, включая расширение тренировочных корпусов и интеграцию специализированных морфологических признаков.

ЗАКЛЮЧЕНИЕ

Проведённое исследование позволило впервые осуществить комплексную сравнительную оценку трёх архитектурно различных систем автоматического распознавания казахской речи — Whisper, Wav2Vec2 и Vosk — на специально созданном авторском речевом корпусе, сбалансированном по полу, возрасту и длине высказывания. Полученные результаты демонстрируют устойчивую межмодельную асимметрию: гибридная система Vosk показала наименьший уровень ошибок и высокую стабильность across-speakers; Wav2Vec2 продемонстрировала умеренное качество и относительно сжатые распределения ошибок; мультиязычная трансформерная система Whisper оказалась наименее адаптированной к агглютинативной морфологии казахского языка, характеризуясь высокой медианной ошибкой и значительной вариативностью. Эти результаты согласуются с выводами исследований, посвящённых влиянию типологических особенностей тюркских языков на устойчивость ASR-моделей (Abudouwaili et al., 2023; Sak et al., 2010; Makhambetov et al., 2015).

Стратифицированный анализ показал, что детская речь, мужские голоса и длинные синтаксически сложные фразы приводят к систематическому увеличению ошибок для всех моделей. Наибольшая деградация наблюдается у Whisper, тогда как Vosk демонстрирует минимальную чувствительность к указанным факторам. Детализированный анализ символных, морфемных и лексических ошибок выявил закономерные подстановки (ң→н, і→ы, қ→к), редукцию морфемных окончаний и трансформацию аффиксальных структур — типичные проблемы ASR-систем при работе с агглютинативными языками (Mussakhojayeva et al., 2023; Karabaliyev & Kolesnikova, 2024).

С практической точки зрения исследование подчёркивает необходимость разработки специализированных казахскоязычных моделей ASR с учётом морфологической структуры языка, расширения корпусов детской речи, а также включения морфологических признаков в архитектуры современных систем распознавания. Кроме того, полученные результаты могут использоваться при создании образовательных, медицинских и государственных сервисов, требующих повышенной точности распознавания речи в условиях низкоресурсных языков.

Работа открывает перспективы для дальнейших исследований, включая разработку морфологически информированных моделей, интеграцию казахского ASR в мультиязычные трансформерные архитектуры, а также построение адаптивных моделей, устойчивых к возрастным и темпово-артикуляционным вариациям. Дальнейшее расширение корпуса, включающее диалекты, шумовые условия и спонтанную речь, позволит повысить обобщающую способность систем и приблизить качество распознавания к уровню высокоресурсных языков.

КОНФЛИКТ ИНТЕРЕСОВ: Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

Abudouwaili, G., Ablez, W., Abiderexiti, K., Wumaier, A., & Yi, N. (2023). Strategies to improve low-resource agglutinative languages morphological inflection. In J. Jiang, D. Reitter, & S. Deng (Eds.), Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL 2023) (pp. 508–520). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.34

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2006.11477

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Conneau, A., Baevski, A., Zhang, Y., et al. (2021). Unsupervised cross-lingual speech representation learning for speech recognition. ICASSP 2021. https://doi.org/10.1109/ICASSP39728.2021.9415029

Coulson, S., & Ma, J. (2022). Morphophonological variation in Turkic languages: Implications for speech technology. Journal of Phonetics, 92, 101114. https://doi.org/10.1016/j.wocn.2022.101114

Karabaliyev, Y., & Kolesnikova, K. (2024). Kazakh speech and recognition methods: Error analysis and improvement prospects. Scientific Journal of Astana IT University, 20, 62–75. https://doi.org/10.37943/20DZGH8448

Khassanov, Y., Mussakhojayeva, S., Mirzakhmetov, A., et al. (2021). A crowdsourced open-source Kazakh speech corpus and initial ASR baseline. EACL 2021, 697–706. https://doi.org/10.18653/v1/2021.eacl-main.58

Klemen, M., Krsnik, L., & Robnik-Šikonja, M. (2023). Enhancing deep neural networks with morphological information. Natural Language Engineering, 29(2), 360–385. https://doi.org/10.1017/S1351324922000080

Kose, H., Saraclar, M., & Ciloglu, T. (2020). Cross-lingual transfer for Turkic automatic speech recognition. Speech Communication, 121, 1–11. https://doi.org/10.1016/j.specom.2020.04.002

Makhambetov, O., Makazhanov, A., Sabyrgaliyev, I., & Yessenbayev, Z. (2015). Data-driven morphological analysis and disambiguation for Kazakh. In CICLing 2015, LNCS 9041 (pp. 151–163). https://doi.org/10.1007/978-3-319-18111-0_12

Mamyrbayev, O., Oralbekova, D., Alimhan, K., Turdalykyzy, T., & Othman, M. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. Scientific Reports, 12, 8337. https://doi.org/10.1038/s41598-022-12260-y

Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., et al. (2019). Automatic recognition of Kazakh speech using deep neural networks. In SPECOM 2019, LNCS 11657 (pp. 465–474). https://doi.org/10.1007/978-3-030-14802-7_40

Mussakhojayeva, S., Khassanov, Y., & Varol, H. A. (2022). KSC2: An industrial-scale open-source Kazakh speech corpus. Interspeech 2022, 1367–1371. https://doi.org/10.21437/Interspeech.2022-421

Mussakhojayeva, S., Khassanov, Y., & Varol, H. A. (2023). Multilingual speech recognition for Turkic languages. Information, 14(2), 74. https://doi.org/10.3390/info14020074

Povey, D., Ghoshal, A., Boulianne, G., et al. (2011). The Kaldi speech recognition toolkit. ASRU 2011. https://doi.org/10.1109/ASRU.2011.6163930

Radford, A., Kim, J. W., Xu, T., Brock, A., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. OpenAI Technical Report. https://doi.org/10.48550/arXiv.2212.04356

Sak, H., Güngör, T., & Saraçlar, M. (2010). Morphology-based and sub-word language modeling for Turkish automatic speech recognition. ICASSP 2010, 5494–5497. https://doi.org/10.1109/ICASSP.2010.5494927

Shivakumar, P. G., & Georgiou, P. (2020). Transfer learning from adult to children speech for automatic speech recognition. Computer Speech & Language, 61, 101077. https://doi.org/10.1016/j.csl.2020.101077

Авторлар туралы мәліметтер
Информация об авторах
Information about authors

Оспанов Манат Габдракипович – докторант, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай қ., Қазақстан
Оспанов Манат Габдракипович – докторант, Костанайский региональный университет имени Ахмета Байтұрсынұлы, г. Костанай, Казахстан
Ospanov Manat Gabdrakipovich – Doctoral Student, Kostanay Regional University named after Akhmet Baitursynuly, Kostanay, Kazakhstan
e-mail: manatog@gmail.com
ORCID: https://orcid.org/0009-0004-5910-0409

Мауленов Қалыбек Сапарұлы – PhD докторы, цифрлық технологиялар және жасанды интеллект бөлімінің бастығы, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай қ., Қазақстан
Мауленов Калыбек Сапарулы – доктор PhD, начальник отдела цифровых технологий и искусственного интеллекта, Костанайский региональный университет имени Ахмета Байтұрсынұлы, г. Костанай, Казахстан
Maulenov Kalybek Saparuly – Doctor of PhD, Head of the Department of Digital Technologies and Artificial Intelligence, Akhmet Baitursynuly Kostanay Regional University, Kostanay, Kazakhstan
e-mail: k.maulenov070693@gmail.com
ORCID: https://orcid.org/0000-0003-4147-3843

Байманқұлов Абдыкәрім Тұңғышбайұлы – физика-математика ғылымдарының докторы, профессор, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай қ., Қазақстан
Байманкулов Абдыкарим Тунгушбаевич – доктор физико-математических наук, профессор, Костанайский региональный университет имени Ахмета Байтұрсынұлы, г. Костанай, Казахстан
Baimankulov Abdykarim Tungushbayevich – Doctor of Physical and Mathematical Sciences, Professor, Kostanay Regional University named after Akhmet Baitursynuly, Kostanay, Kazakhstan.
e-mail: bat_56@mail.ru
ORCID: https://orcid.org/0000-0002-6435-9560