МОДЕЛИ РАСПОЗНАНИЯ И СИНТЕЗА РЕЧИ

Рагачурина Нозима Акмаловна
Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, PhD,

Annotatsiya: sun'iy intellekt va mashinani o'rganish texnologiyalarining jadal rivojlanishi bilan nutqni avtomatik aniqlash (ASR) va matnni nutqqa aylantirish (TTS) jamiyatning raqamli transformatsiyasining asosiy tarkibiy qismlariga aylandi. Turkiy tillar oilasidan bo'lgan o'zbek tili cheklangan til modellari va yuqori sifatli nutq sintezi tizimlariga ega bo'lgan resurslar bilan cheklangan til bo'lib qolmoqda. Alohida e'tiborlisi, uning agglyutinativ tuzilishi, boy unlilar tizimi va fonematik o'zgaruvchanligi kabi lingvistik va texnik to'siqlarga qaratiladi. Nutqni aniqlash tizimlari BLEU, WER, TER, chrF va COMET yordamida baholanadi, nutq sintezi esa MCD, PESQ, STOI va DNSMOS yordamida baholanadi, bu esa leksik-semantik va akustik-idrok xususiyatlarini qamrab oldadi. Nutqni matnga aylantirish (STT) uchun Soyle domenga xos ma'lumotlar bo'yicha eng yaxshi natijalarni ko'rsatdi, Voiser esa muvozanatli aniqlikni namoyish etdi va GPT-4 Transcribe kuchli semantik saqlanishni ta'minladi. Aksincha, Whisper eng yomon natijalarni ko'rsatdi, bu esa o'zbek tili uchun qo'shimcha moslashishni talab qildi.

Kalit so‘zlar: o'zbek tili, nutqni avtomatik aniqlash, nutqni matnga aylantirish, fonematik

Аннотация: в условиях стремительного развития технологий искусственного интеллекта и машинного обучения, автоматическое распознавание речи (ASR) и преобразование текста в речь (TTS) стали ключевыми компонентами цифровой трансформации общества. Узбекский язык, как представитель тюркской языковой семьи, остается языком с ограниченными ресурсами, имеющим ограниченные языковые модели и высококачественные системы синтеза речи. Особое внимание уделяется лингвистическим и техническим барьерам, включая агглютинативную структуру, богатую систему гласных, и фонемную изменчивость. Системы распознавания речи оценивались с использованием BLEU, WER, TER, chrF и COMET, а синтез речи — с помощью MCD, PESQ, STOI и DNSMOS, таким образом, охватывая как лексико-семантические, так и акустико-перцептивные характеристики. Для преобразования речи в текст (STT) наилучшие результаты показала система Soyle на данных, специфичных для данной области, в то время как Voiser показала сбалансированную точность, а GPT-4 Transcribe обеспечила надежное сохранение семантики. В отличие от этого, Whisper показала наихудшие, требуя дальнейшей адаптации для узбекского языка.

ISSN 3030-3907

Ключевые слова: узбекский язык, автоматическое распознание речи, преобразование речи в текст, фонетика

Abstract: with the rapid development of artificial intelligence and machine learning technologies, automatic speech recognition (ASR) and text-to-speech (TTS) have become key components of the digital transformation of society. Uzbek, a member of the Turkic language family, remains a resource-constrained language with limited language models and high-quality speech synthesis systems. Particular attention is paid to linguistic and technical barriers, including its agglutinative structure, rich vowel system, and phonemic variability. Speech recognition systems were evaluated using BLEU, WER, TER, chrF, and COMET, while speech synthesis was assessed using MCD, PESQ, STOI, and DNSMOS, thus encompassing both lexical-semantic and acoustic-perceptual characteristics. For speech-to-text (STT), Soyle performed best on domain-specific data, while Voiser demonstrated balanced accuracy, and GPT-4 Transcribe ensured robust semantic preservation. In contrast, Whisper performed the worst, requiring further adaptation for the Uzbek language.

Keywords: uzbek language, automatic speech recognition, speech-to-text, phonetic

Введение

Внедрение систем автоматического распознавания речи (ASR) открывает значительные возможности для различных аспектов общества, включая повышение производительности и расширение возможностей для пожилых людей и людей с ограниченными возможностями. Однако эффективное и безопасное использование этих технологий требует высокой точности и способности модели речи понимать язык в контекстно-зависимой форме.

Системы ASR представляют собой набор методов и алгоритмов, направленных на преобразование устной речи в текстовую форму. В последние десятилетия ASR эволюционировала из экспериментальной технологии в неотъемлемую часть повседневной жизни. Современные системы ASR широко используются в различных областях, включая голосовых помощников, системы автоматического перевода, управление интеллектуальными устройствами и обработку аудиоданных.

Автоматическое распознавание речи для узбекского языка представляет собой ряд существенных проблем, как технических, так и лингвистических. Несмотря на растущий интерес к изучению и оцифровке узбекского языка, он остается языком с ограниченными ресурсами, особенно в контексте речевых технологий.

ISSN 3030-3907

Тюркские языки, такие как узбекский, киргизский, казахский и татарский, считаются языками с ограниченными ресурсами, и все они являются агглютинативными. В агглютинативном языке слова образуются путем сочетания морфем и аффиксов. узбекский язык имеет сложную морфологию. В узбекском языке, как и в большинстве тюркских языков, слова образуются путем добавления аффиксов к корню. Эти аффиксы могут передавать одно грамматическое или словообразовательное значение, что является существенным в этом процессе. Аффиксы не сливаются и сохраняют свою индивидуальность. В работах [1] авторы объясняют, как образуются слова в тюркских языках. Они отмечают, что все тюркские языки используют четыре типа аффиксов: множественного числа, личные, падежные и притяжательные аффиксы. Упомянутые работы подчеркивают, что тюркские языки следуют универсальному принципу словообразования, основанному на этих четырех основных категориях. Порядок этих аффиксов не случаен; он регулируется строгими морфологическими и фонологическими правилами, особенно гармонией гласных, что обеспечивает фонетическую согласованность и грамматическую связность. За исключением турецкого, большинство тюркских языков имеют ограниченное количество доступных и размеченных аудио- и текстовых данных, что усложняет разработку алгоритмов распознавания речи и других приложений речевых технологий. Кроме того, высокая морфологическая сложность и агглютинативная природа этих языков создают дополнительные проблемы при разработке моделей автоматического распознавания речи.

Современные исследования в области автоматического распознавания речи неизбежно сталкиваются с рядом существенных проблем, включая ограниченную доступность и непостоянное качество аудиозаписей на большинстве тюркских языков. Среди них турецкий язык выделяется относительно большими и качественными корпусами, в то время как узбекский, татарский и казахский языки остаются сравнительно малоизученными, с меньшими наборами данных и менее стабильными условиями записи. Дополнительные проблемы включают высокую вариативность речевых паттернов, необходимость соблюдения строгих правил защиты данных. Хотя этап сбора аудиоданных является важным компонентом в разработке речевых технологий, гораздо более сложной задачей является обеспечение высокого качества и согласованности как самих аудиофрагментов, так и сопровождающей их аннотации.

Аудиоданные часто содержат элементы, позволяющие идентифицировать говорящего, что требует получения явного и документально оформленного согласия участников на использование их

ISSN 3030-3907

голосовой информации в исследовательских или коммерческих целях. В исследовании использовались открытые данные из интернет-источников, которые не нарушают права на неприкосновенность частной жизни говорящих.

Процесс аннотирования аудиоматериалов неизбежно требует текстовой транскрипции, качество которой напрямую влияет на успех обучения модели. Наличие неполной, неточной или субъективной транскрипции, а также использование нестандартизированных сокращений и обозначений значительно снижает эффективность последующей обработки.

Отдельной трудностью является техническое качество аудиозаписей. Наличие фонового шума, акустических искажений, низкой разборчивости речи или плохих условий записи усложняет обучение моделей. Это требует использования специализированных алгоритмов предварительной обработки сигнала, включая шумоподавление, нормализацию громкости и методы фильтрации. В некоторых случаях данные, которые невозможно исправить, исключаются из обучающего набора, что увеличивает стоимость ресурсов и времени, необходимых для формирования всесторонней обучающей выборки.

Один из практических подходов к компенсации недостатка аудиоданных — создание корпусов синтетической речи, которые генерируют аудиосигналы на основе технологий преобразования текста в речь (TTS). Этот метод позволяет масштабируемо получать аудиоматериалы с контролируемыми параметрами, такими как скорость речи, интонация, акустические условия и характеристики речи говорящего, обеспечивая формирование сбалансированных и разнообразных обучающих выборок. Альтернативный, широко используемый подход — извлечение аудиоданных из открытых интернет-источников, включая подкасты, видеохостинги и общедоступные базы данных речи. Кроме того, данные могут быть собраны путем целенаправленной записи речи пользователя с использованием специализированного оборудования и программного обеспечения. Эти методы предоставляют реалистичные примеры спонтанной и формальной речи на узбекском языке в различных акустических и лингвистических контекстах, что имеет решающее значение для тестирования и обучения моделей.

Большие наборы данных, передовые нейронные архитектуры и мощные вычислительные ресурсы стали ключевыми факторами для разработки надежных систем автоматического распознавания речи (ASR) и синтеза речи (TTS). Текущие исследования сосредоточены на повышении устойчивости моделей к шуму и их способности адаптироваться к языкам с

ISSN 3030-3907

ограниченными ресурсами, улучшении способности к обобщению между языками и повышении естественности и выразительности синтезированной речи.

Технологии ASR и TTS значительно улучшились за последние годы благодаря достижениям в области глубокого обучения, увеличению доступных речевых корпусов и увеличению вычислительной мощности. Эти достижения значительно повысили точность и эффективность существующих моделей для многих языков, особенно для языков с большими ресурсами. Однако существует категория языков с ограниченными ресурсами, для которых разработка высококачественных систем ASR и TTS остается сложной задачей. Ограниченное количество данных для таких языков, как языки меньшинств или региональные диалекты, накладывает существенные ограничения на разработку соответствующих технологий.

Быстрое развитие автоматического распознавания речи было возможным благодаря достижениям в области глубокого обучения. Потенциал синтеза речи для расширения обучающих данных ASR продемонстрирован в [2], где синтезированная речь улучшила точность, хотя разрыв с естественной речью остается. Многоязычная сквозная система ASR с интегрированным прогнозированием LID в рамках RNN-T была предложена в [3], достигнув точности LID 96,2% и конкурентоспособного WER. Универсальная речевая модель (USM), охватывающая более 100 языков, была представлена в [4] как крупномасштабная многоязычная базовая модель для распознавания речи. Модель Whisper также привлекла значительное внимание благодаря своей высокой межъязыковой производительности. Стратегии тонкой настройки для Whisper в условиях ограниченных ресурсов, включая узбекский язык, были систематически оценены в [5], показав, что соответствующая адаптация значительно повышает точность распознавания при одновременном снижении вычислительных затрат.

Рассмотренные исследования были отобраны для обеспечения всестороннего понимания последних достижений, методологий и проблем в разработке речевых и языковых технологий для тюркских языков. Анализируя работы, связанные с автоматическим распознаванием речи (ASR), синтезом речи (TTS), языковым моделированием и многоязычной обработкой — особенно для таких языков, как узбекский, казахский, турецкий и другие, — мы стремимся выявить эффективные стратегии, переносимые методы и пробелы, которые остаются нерешенными. Все рассмотренные в этом разделе исследования изучали различные аспекты

ISSN 3030-3907

автоматического распознавания речи, синтеза речи и перевода речи для тюркских языков. Результаты данного исследования, анализирующего существующие модели, служат основой для предлагаемой методологии, ориентированной на повышение производительности сквозных систем ASR и TTS, улучшение языкового представления в многоязычных моделях и разработку надежных, масштабируемых инструментов для тюркских языков с минимальным количеством аннотированных ресурсов. Этот подход играет важную роль в развитии узбекских языковых технологий, и в исследовании особое внимание уделяется его практическому применению.

СПИСОК ЛИТЕРАТУРЫ

1. Tukeyev, U.; Karibayeva, A.; Zhumanov, Z. Morphological Segmentation Method for Turkic Language Neural Machine Translation. Cogent Eng. 2020, 7, 1832403. [CrossRef]

2. Rosenberg, A.; Zhang, Y.; Ramabhadran, B.; Jia, Y.; Moreno, P.; Wu, Y.; Wu, Z. Speech recognition with augmented synthesized speech. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop, Singapore, 14–18 December 2019;

3. Zhang, C.; Li, B.; Sainath, T.; Strohman, T.; Mavandadi, S.; Chang, S.-Y.; Haghani, P. Streaming end-to-end multilingual speech recognition with joint language identification. In Proceedings of the INTERSPEECH, 2022, Incheon, Korea, 18–22 September 2022;

4. Zhang, Y.; Han, W.; Qin, J.; Wang, Y.; Bapna, A.; Chen, Z.; Chen, N.; Li, B.; Axelrod, V.; Wang, G.; et al. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv* 2023;

5. Liu, Y.; Yang, X.; Qu, D. Exploration of Whisper Fine-Tuning Strategies for Low-Resource ASR. *EURASIP J. Audio Speech Music Process.* 2024;

ISSN 3030-3907