ПЛАТФОРМА ДЛЯ УЗБЕКСКОГО ASR И TTS С МОДУЛЬНОЙ АРХИТЕКТУРОЙ И ОТКРЫТЫМИ КОРПУСАМИ

Авезов Сухроб Собирович

PhD, преподаватель кафедры русского языка и литературы

Бухарский государственный университет

senigama1990@mail.ru

Аннотация: *Исследование демонстрирует революционный прогресс узбекских речевых технологий с 2020 года. Международное сотрудничество между Казахстаном и Узбекистаном создало открытые корпусы и достигло современных показателей: 14.3% WER для ASR и 4.36 MOS для TTS. Агглютинативная природа языка и двойной алфавит остаются ключевыми вызовами для нейронных архитектур.*

Ключевые слова: *автоматическое распознавание речи (ASR), синтез речи (TTS), узбекский язык, трансформеры, USC корпус, common voice, агглютинативные языки.*

Исследование узбекских технологий автоматического распознавания речи (ASR) и синтеза речи (TTS) показывает быстрое развитие научно-исследовательской экосистемы с 2020 года, достигнувшей современных показателей качества 14.3% WER для ASR и 4.36 MOS для TTS. Это представляет значительный прорыв для языка с низким ресурсообеспечением, где международное сотрудничество между Казахстаном и Узбекистаном создало первые открытые стандартизированные корпусы и модели мирового уровня.

Особое значение имеет то, что узбекский язык преодолел статус языка с критически низкими ресурсами благодаря созданию USC корпуса (105 часов) и активному участию в проекте Mozilla Common Voice (265 часов). Современные подходы на основе трансформеров и мультиязычного обучения

демонстрируют превосходство над традиционными статистическими методами на 20-30%. Агглютинативная природа узбекского языка и проблемы с двойным алфавитом (латиница/кириллица) остаются ключевыми техническими вызовами, требующими специализированных архитектурных решений.

Основной центр исследований сосредоточен вокруг сотрудничества между Институтом интеллектуальных систем и искусственного интеллекта (ISSAI) Назарбаев университета в Казахстане и Ташкентским университетом информационных технологий (TUIT). Эта коллаборация под руководством Ерболата Хасанова (ISSAI) и Мухаммаджона Мусаева (TUIT) создала фундаментальную исследовательскую базу для узбекских речевых технологий.

Ключевые исследователи включают Саиду Мусахожаеву (представляла результаты на INTERSPEECH 2022), Ильёса Худжаёрова (соавтор множества публикаций USC), и Хусейна Атакана Варола (старший научный руководитель проектов). Исследовательская группа также включает А.Мухамадиева (TUIT, отдел ИИ) и О.Джураева (TUIT, отдел аппаратно-программных систем управления), создавших первые глубокие нейронные модели для узбекского языка.

Международная природа исследований подчеркивается корейским финансированием через J.Cho и европейскими связями через SPECOM и другие международные конференции. Эта модель сотрудничества стала образцом для развития речевых технологий в Центральной Азии.

Технические показатели узбекского ASR продемонстрировали драматические улучшения с 2020 по 2024 год. Первоначальная работа USC корпуса 2021 года показала 17.4% WER, что уже было значительным достижением для языка без предыдущих открытых ресурсов. Однако к 2022 году исследователи TUIT достигли прорывного результата 14.3% WER используя архитектуру End-to-End Transformer с CTC+Attention на расширенном датасете 207 часов.

Сравнительный анализ различных архитектур показывает превосходство современных подходов: E2E-T (CTC+Attention) достигает 14.3% WER, в то время как традиционные DNN-HMM системы показывают только 19.4% WER. Символьные ошибки (CER) также демонстрируют впечатляющие результаты - 5.26% CER с нейронными языковыми моделями против 5.41% с ограниченным словарем.

Особенно впечатляющими являются результаты для распознавания команд: 97.96% точность с использованием Audio Spectrogram Transformer (AST) на датасете из 28 спикеров, что демонстрирует практическую применимость технологии в реальных сценариях.

Техническая экосистема узбекского ASR/TTS характеризуется модульным дизайном с гибкой интеграцией современных компонентов. Ведущие фреймворки включают ESPnet для end-to-end обработки речи, NVIDIA NeMo для масштабируемого ИИ, и специализированные решения как TurkicTTS для мультиязычного синтеза.

Transformer-архитектуры стали доминирующим подходом с конкретными спецификациями: 12 слоев кодировщика, 6 слоев декодера, 8 головок внимания на слой, и 80-мерные log-Mel спектрограммы как входные данные. CTC вес 0.3 для совместного обучения и 0.6 для совместного декодирования оказались оптимальными параметрами.

Wav2Vec2 адаптации показали значительный потенциал с моделями как lucio/xls-r-uzbek-cv8 (fine-tuned на Common Voice 8.0) и sarahai/uzbek-stt-3 (специализированный на юридических и военных датасетах). Эти модели достигают WER 0.2588 используя 208,000+ аудиофайлов, демонстрируя эффективность трансферного обучения для языков с ограниченными ресурсами.

Whisper интеграция представляет более сложный случай: несмотря на официальную поддержку узбекского языка в 96-языковой модели, практическое fine-tuning сталкивается с проблемами токенизации и деградацией мультиязычной производительности.

USC (Uzbek Speech Corpus) остается краеугольным камнем узбекских речевых исследований с 105 часами тщательно проверенного аудио от 958 спикеров. Датасет организован с 80% тренировочной выборки, 10% валидации, и 10% тестирования, обеспечивая стандартизированные бенчмарки. Creative Commons Attribution 4.0 лицензия делает корпус доступным как для академических, так и для коммерческих приложений через HuggingFace и GitHub.

FeruzaSpeech представляет уникальную 60-часовую коллекцию высококачественных студийных записей одного женского диктора с дуальными транскрипциями (латиница и кириллица). Этот корпус особенно ценен для TTS приложений, достигая впечатляющих 4.05% WER на тестовой выборке и демонстрируя важность качества данных над количеством.

Mozilla Common Voice обеспечивает массовые краудсорсинговые данные с 265 часами валидированного аудио от 2000+ участников. Это представляет крупнейший открытый ресурс узбекской речи и демонстрирует успешное международное сотрудничество в создании языковых ресурсов.

Коммерческие ресурсы включают DataoceanAI корпус с 392 часами профессионально записанного контента от 200 спикеров, покрывающий 11 индустрий и достигающий 95% точности на уровне предложений. UzWordNet предоставляет лексико-семантическую базу данных с 28,140 синсетами, совместимую с Princeton WordNet.

Агглютинативная природа узбекского языка создает фундаментальные технические проблемы для речевых систем. В то время как английский содержит десятки тысяч слов, узбекский может генерировать миллионы потенциальных словоформ через комбинации суффиксов, приводя к критической разреженности данных в статистических моделях и объясняя превосходство end-to-end нейронных подходов.

Mozilla Common Voice служит основной платформой международного сотрудничества с сильным узбекским сообществом, показавшим «удивительный рост» в последних релизах. Проект UzbekVoice.ai,

возглавляемый Мухаммадом Амином Кодировым, собрал около 1,400 часов высококачественного аудио и стремится внести свой датасет в Common Voice платформу с соответствующей атрибуцией.

Microsoft Azure Cognitive Services официально поддерживает узбекский язык в Neural TTS с 2022 года, когда узбекский был добавлен к 49 новым языкам/вариантам, доведя общее количество поддерживаемых языков до 119+. Facebook/Meta MMS проект включает узбекские модели TTS (facebook/mms-tts-uzb-script_cyrillic) с поддержкой кириллицы через архитектуру VITS.

NVIDIA NeMo framework интегрирует узбекские данные Common Voice в multilingual ASR обучение, предоставляя бесплатный доступ к моделям через партнерство с Mozilla. Несколько GitHub репозиториев используют NeMo для разработки узбекского ASR.

Современные системы узбекского ASR достигают состояния искусства 14.3% WER с E2E Transformer архитектурами, что сопоставимо с performance других языков среднего ресурсообеспечения. Character Error Rate показывает еще более впечатляющие результаты с 5.26% CER для нейронных языковых моделей, демонстрируя особую эффективность символьного уровня обработки для агглютинативных языков.

TTS системы показывают выдающееся качество с 4.36/5.0 MOS scores для специализированных узбекских систем на основе Tacotron+WaveGAN. Мультиязычная TurkicTTS система демонстрирует более скромные результаты (2.85/5.0 качество, 80% понятность, 45% разборчивость), что отражает компромиссы между мультиязычным покрытием и языкоспецифическим качеством.

Распознавание команд достигает 97.96% точности с Audio Spectrogram Transformer подходами, указывая на высокий потенциал для практических приложений в голосовых интерфейсах и умных системах.

Ландшафт узбекских ASR и TTS технологий трансформировался от критически низкоресурсного языка до устойчивой исследовательской

экосистемы с международным признанием за пятилетний период 2020-2025. Ключевые достижения включают: создание стандартизированных корпусов 105+ часов, достижение state-of-the-art ASR performance 14.3% WER, функциональные TTS системы с 4.36/5.0 MOS scores, и open-source доступность, обеспечивающая участие более широкого исследовательского сообщества.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Elov B., Xudayberganov N. O ‘zbek tili korpusi matnlarini pos teglash usullari //Computer Linguistics: problems, solutions, prospects. – 2024. – Т. 1. – №. 1.

2. Qarshiyev A., Tursunov M., Maxmidov S. O ‘zbek tili milliy korpusini loyihalash //Computer linguistics: problems, solutions, prospects. – 2022. – Т. 1. – №. 1.

3. Elov B., Abdullayeva O. O ‘zbek tili korpusini sintaktik teglash masalasi //Computer Linguistics: problems, solutions, prospects. – 2024. – Т. 1. – №. 1.

4. Sobirovich S. A. A PRAGMATICALLY ORIENTED APPROACH TO GENERATIVE LINGUISTICS //CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES. – 2024. – Т. 5. – №. 04. – С. 69-75.

5. Авезов С. КОРПУСНАЯ ЛИНГВИСТИКА: НОВЫЕ ПОДХОДЫ К АНАЛИЗУ ЯЗЫКА И ИХ ПРИЛОЖЕНИЯ В ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ //International Bulletin of Applied Science and Technology. – 2023. – Т. 3. – №. 7. – С. 177-181.