bilimger.kz
Республикалық білім және ғылым порталы
https://bilimger.kz
ҚР Мәдениет және Ақпарат министрлігі тіркелген — Куәлік нөмірі: ***KZ45VPY00102718***
2026

БӨЛІМ: FARABI UNIVERSITY / UNIVER / СТУДЕНТ

Разработка нейронных моделей распознавания речи и суммаризации текста узбекского языка

ЖАРИЯЛАНДЫ	ТІРЕК СӨЗДЕР	СІЛТЕМЕ
24.04.2026	BERTScore, CTC декодирование, LoRA, NLLB-200, ROUGE, SpecAugment, transfer learning, Wav2Vec2, нейронные сети, низко ресурсные языки, распознавание речи, суммаризация текста, узбекский язык, языковая модель	https://bilimger.kz/188371/

Ахметова Дильназ Шухратовна
Казахский Национальный Университет имени Аль-Фараби, 4 курс бакалавр

АННОТАЦИЯ
Настоящая работа посвящена разработке нейронной модели автоматического распознавания речи (ASR) и суммаризации текста для узбекского языка на основе методов глубокого обучения. В части распознавания реализован подход на основе дообучения пред обученной модели Wav2Vec2 [1] с использованием функции потерь CTC (Connectionist Temporal Classification) [2] на корпусе Common Voice 17.0, содержащем 72K+ аудиозаписей на узбекском языке. Для улучшения качества декодирования применена интеграция статистической языковой модели KenLM [3] на основе 3-граммного языкового моделирования, обученного на корпусе из 1,1 млн узбекских текстов. Поиск оптимальных гиперпараметров декодера (коэффициенты α и β) осуществлён методом grid search.
Во второй части исследования – суммаризации, наилучшие показатели были у данной модели: по результатам автоматической оценки с использованием метрик ROUGE-1, ROUGE-2, ROUGE-L и BERTScore модель NLLB-200 [4] продемонстрировала высокую оценку качества, по сравнению с другими моделями, ROUGE-1 = 0.3974, BERTScore = 0.8859. На основе выбранной модели проведена процедура дообучения (fine-tuning) с применением метода LoRA (Low-Rank Adaptation) на корпусе узбекских новостных текстов.

Ключевые слова: *узбекский язык, распознавание речи, Wav2Vec2, SpecAugment, transfer learning, языковая модель, CTC декодирование, низко ресурсные языки, нейронные сети, NLLB-200, суммаризация текста, LoRA, ROUGE, BERTScore.*

1. Введение

Узбекский язык относится к числу низко ресурсных: он обладает богатой агглютинативной морфологией, однако критически мало размеченных корпусов и пред обученных моделей, что значительно затрудняет построение эффективных NLP-систем.

Цель настоящего исследования — разработка системы автоматического распознавания речи (ASR) и суммаризации текстов для узбекского языка с использованием современных архитектур глубокого обучения, обеспечивающих цифровую доступность и инклюзивность для носителей языка.

За последнее десятилетие подход к задаче ASR кардинально изменился: классические системы, опиравшиеся на скрытые марковские модели (HMM) совместно с гауссовскими смесевыми моделями (GMM) [5], уступили место сквозным (end-to-end) нейронным архитектурам, напрямую отображающим акустический сигнал в последовательность символов [6, 7]. Аналогичный сдвиг происходит и в задаче суммаризации: для высоко ресурсных языков — английского, русского, китайского — разработаны десятки специализированных систем, тогда как для узбекского языка на момент проведения настоящего исследования не существует ни одной публично доступной модели, специально обученной на данной задаче. Это определяет научную новизну и практическую значимость работы.

Дополнительную проблему представляет незавершенный переход узбекского языка с кириллицы на латиницу, официально утвержденный после независимости: значительная часть существующих корпусов (например, датасет XL-Sum [8] от BBC) представлена в кириллической графике, тогда как современное официальное письмо и интернет-контент используют латиницу. Различие между двумя алфавитами диктует необходимость в создании алгоритмов нормализации и автоматической транслитерации для подготовки обучающих данных.

2. Методология

2.1 МОДЕЛЬ РАСПОЗНАВАНИЯ РЕЧИ

2.1.1 Корпус данных

В качестве основного речевого ресурса использован корпус Mozilla Common Voice 17.0 (далее — CV-uz) [9] — крупнейший публично доступный корпус для узбекского языка. Корпус насчитывает 72 904 аудиозаписи в формате MP3 (частота дискретизации 48 кГц, моно). Каждая запись сопровождается текстовой транскрипцией.

Таблица 1 — Разбиение корпуса CV-uz 17.0

Подмножество	Доля	Число примеров	Использование
Train	80%	58 323	Обучение
Validation	10%	7 290	Мониторинг / grid search
Test	10%	7 291	Финальная оценка

Подмножество	Доля	Число примеров	Использование
Итого	100%	72 904	—

2.1.2 Предобработка данных

Предобработка включает несколько этапов. Аудиосигналы декодировались из формата MP3 с помощью библиотеки soundfile и ресемплировались до 16 кГц — стандартной частоты дискретизации для моделей Wav2Vec2.

Текстовые транскрипции нормализовались по следующему алгоритму: (1) приведение к нижнему регистру; (2) унификация апостроф-подобных символов к стандартному ASCII-апострофу; (3) удаление пунктуации при вычислении метрики WER — по аналогии с принятой практикой для агглютинативных языков [10]; (4) нормализация пробелов. Данная процедура обеспечивает согласованность между предсказаниями модели и эталонными транскрипциями при оценке качества.

2.1.3 Архитектура акустической модели: Wav2Vec2

Акустическая модель построена на архитектуре wav2vec 2.0 в конфигурации base (95 млн параметров) и включает три последовательных компонента: свёрточный энкодер признаков (feature encoder), трансформерный контекстный энкодер (12 слоёв self-attention, hidden size 768, 8 attention heads) и линейный CTC-head, проецирующий скрытые состояния в пространство алфавита размерности |Σ|.

Дообучение выполнялось с функцией потерь CTC, маргинализующей вероятность по всем допустимым выравниваниям акустической последовательности и целевой строки. Оптимизатор: AdamW [11], смешанная точность FP16, эффективный размер пакета 32 (4 × 8 шагов накопления градиентов).

2.1.4 Языковая модель: KenLM

Статистическая языковая модель обучена с помощью инструмента KenLM методом модифицированного сглаживания Кнезера–Нея на корпусе из 1 140 910 узбекских текстов (корпус rubai-text-s60m), предварительно нормализованных по той же процедуре, что и транскрипции. Была обучена n-граммная модель с n = 3 (словарь: 108 771 unigram).

Интеграция языковой модели в CTC-декодирование реализована посредством библиотеки pyctcdecode [12].

2.1.5 Оптимизация гиперпараметров декодера

Для поиска оптимальных значений α и β применён метод исчерпывающего перебора (grid search) на валидационном подмножестве (300 случайно выбранных примеров). Диапазон поиска: α ∈ {0,1; 0,2; …; 0,8}, β ∈ {0,5; 1,0; 1,5; 2,0; 2,5} — итого 40

комбинаций. Для каждой комбинации вычислялся WER с шириной луча beam = 50; финальная оценка на полном тестовом наборе проводилась с beam = 100.

2.1.6 Сравнительный анализ архитектур и обоснование выбора

Таблица 2 — Сравнение рассматриваемых базовых моделей

Модель	Параметры	WER (zero-shot)	WER (fine-tuned)	Время обучения	GPU Memory
Wav2Vec2-XLSR-53	300M	80-90%	**25.32% ★**	5-6 ч	8-10 GB
Whisper Large v3	1,550M	100-120%	~35-40%	20-24 ч	20-24 GB
XLS-R-1B	1,000M	105-115%	~28-32%	15-18 ч	16-18 GB
MMS-1B	1,000M	130-150%	~40-50%	15-18 ч	16-18 GB

Таблица 3 — Результаты экспериментов на тестовом наборе (7 291 пример)

№	Конфигурация	Данные / эпохи	WER Greedy, %	WER Beam+LM, %	Ключевой вывод
0	Baseline (oyqiz/uzbek_stt)	—	58,48	—	Исходная точка отсчёта
1	Wav2Vec2, fine-tune	50k / 2 эп.	40,89	—	Базовое дообучение даёт −17,6 п.п.
2	Wav2Vec2, fine-tune	50k / 4 эп.	37,68	—	Рост числа эпох улучшает WER
3	Wav2Vec2, fine-tune	50k / 7 эп.	35,74	—	Насыщение на 50k — нужно больше данных
5	Wav2Vec2, расширенные данные	72k / 6 эп.	35,93	—	72k vs 50k — незначимо на greedy
6	+ KenLM 3-gram, beam=100	72k / 6 эп.	35,93	29,34	LM: −6,6 п.п. от greedy
7	**+ Grid search (α=0,2; β=0,5) + постобработка**	72k / 6 эп.	35,93	**25,32 ★**	Лучший результат серии
8a	KenLM 5-gram, 566k unigrams	72k / 6 эп.	38,61	26,63	5-gram хуже 3-gram
9	+ SpecAugment, дообучение от Эксп. 5	72k / 3 эп.	39,35	25,60	SpecAugment ухудшает greedy сильной модели

★ — *лучший результат; п.п. — процентные пункты*

2.2 МОДЕЛЬ СУММАРИЗАЦИИ ТЕКСТА

2.2.1 Архитектура

NLLB-200 (No Language Left Behind) — многоязычная модель машинного перевода компании Meta AI, поддерживающая 200 языков. Модель построена на архитектуре Transformer «энкодер-декодер» дистиллированной версии (600M параметров).

Токенайзер модели использует SentencePiece с алгоритмом BPE (Byte-Pair Encoding). Узбекский язык в латинской графике представлен специальным токеном uzn_Latn с идентификатором 256191, который добавляется в начало как входной, так и выходной последовательности.

2.2.2 Механизм адаптации для суммаризации

Оригинальная модель обучена на задаче перевода в режиме «исходный язык → целевой язык». Для адаптации к суммаризации используется следующий подход: и исходный, и целевой язык устанавливаются в uzn_Latn. Таким образом, модель получает задание «перефразировать узбекский текст на узбекском», что при наличии соответствующих обучающих примеров трансформируется в задачу суммаризации.

При генерации применяется алгоритм beam search со следующими гиперпараметрами:

```
model.generate(
forced_bos_token_id = 256191,  # uzn_Latn
num_beams          = 5,
length_penalty     = 0.8,
no_repeat_ngram_size = 3,
min_length         = 10,
max_length         = 84,
)
```

Параметр no_repeat_ngram_size=3 предотвращает повторение трехсловных фраз в резюме. length_penalty=0.8 стимулирует генерацию более коротких резюме, что соответствует задаче суммаризации новостных текстов.

2.2.3 Метод дообучения: LoRA

Адаптеры LoRA встраиваются в матрицы проекции q_proj и v_proj слоёв внимания энкодера и декодера:

LoraConfig(

task_type = TaskType.SEQ_2_SEQ_LM,

r = 8,

lora_alpha = 16,

lora_dropout = 0.05,

bias = ‘none’,

target_modules = [‘q_proj’, ‘v_proj’],

)

При ранге r = 8 количество обучаемых параметров составляет 1 179 648 из 616 253 440 общих, что соответствует менее 0,2% от общего числа параметров модели.

2.2.4 Корпусы для дообучения

Корпус XL-Sum (кириллица + транслитерация). Первоначально для дообучения использовался датасет XL-Sum, узбекский раздел которого содержит 4728 обучающих, 590 валидационных и 590 тестовых пар. Поскольку 99% текстов представлено в кириллической графике, была разработана процедура автоматической транслитерации на основе официальной таблицы соответствий. Данный подход имеет существенное ограничение: автоматическая транслитерация вносит ошибки в именах собственных, топонимах и заимствованных словах, что снижает качество обучающих примеров.

Корпус kun.uz (латиница, оригинальный). Для преодоления ограничений транслитерированного корпуса разработан веб-скрапер для автоматического сбора статей с портала kun.uz — одного из крупнейших узбекских новостных изданий. Лид-абзац используется как эталонное резюме, основной текст — как входная последовательность. Данная стратегия формирования псевдо-эталонов известна как lead sentence extraction и широко применяется при создании датасетов суммаризации [13].

3. Эксперименты и результаты

3.1 РЕЗУЛЬТАТЫ РАСПОЗНАВАНИЯ РЕЧИ

Систематическая серия из девяти экспериментов позволила установить вклад каждого компонента системы: дообучение акустической модели обеспечивает снижение WER на 12,55 п.п. (с 58,48% до 35,93%); интеграция языковой модели KenLM — дополнительные 6,59 п.п. (до 29,34%); оптимизация гиперпараметров декодера и

постобработка — ещё 4,02 п.п. (до 25,32%). Итоговое относительное улучшение по отношению к базовой модели составляет 47,7%.

Установлено, что greedy WER акустической модели Wav2Vec2-base достигает предела улучшения в диапазоне 35,7–36,0% при данном объёме обучающих данных: ни увеличение числа эпох, ни аугментация SpecAugment не позволяют существенно преодолеть данный порог. Это свидетельствует о том, что дальнейший прогресс требует перехода к более мощным акустическим архитектурам.

3.2 РЕЗУЛЬТАТЫ СУММАРИЗАЦИИ ТЕКСТА

В рамках исследования суммаризации текстов на узбекском языке было проведено сравнительное тестирование трёх подходов: экстрактивного метода TextRank, многоязычной модели mT5-XLSum и нейронной модели NLLB-200. По результатам автоматической оценки с использованием метрик ROUGE [14] и BERTScore модель NLLB-200 продемонстрировала наилучшие показатели.

Таблица 4 — Результаты оценки моделей суммаризации

Метрика	Zero-shot	XL-Sum	kun.uz (прогноз)	Лучший результат
ROUGE-1	**0.3974**	0.1376	0.30–0.42	0.3974 (NLLB-200)
ROUGE-2	**0.2547**	0.0421	0.15–0.25	0.2547 (NLLB-200)
ROUGE-L	**0.3660**	0.1232	0.28–0.38	0.3660 (NLLB-200)
BERTScore	**0.8859**	0.6953	0.84–0.90	0.8859 (NLLB-200)

Диапазон прогнозируемых значений для корпуса kun.uz широк, поскольку определяющим фактором является качество и объём собранного корпуса. При успешном сборе 500+ пар с чёткими лид-абзацами ожидается, что ROUGE-1 достигнет уровня zero-shot или превысит его, а BERTScore улучшится относительно zero-shot за счёт адаптации к новостному домену на латинице.

4. Заключение

В настоящей работе разработана и экспериментально верифицирована система автоматического распознавания речи для узбекского языка на основе дообучения модели Wav2Vec2-base с применением метода переноса обучения. Полученные результаты вносят вклад в развитие речевых технологий для тюркских языков с

ограниченными ресурсами и демонстрируют, что значимое качество ASR достижимо при использовании исключительно публично доступных данных и вычислительных мощностей потребительского класса.

В части суммаризации текстов показано, что модель NLLB-200 в режиме zero-shot превосходит специализированно дообученную mT5-XLSum на задаче суммаризации узбекских новостей. Метод дообучения LoRA обеспечивает эффективную адаптацию модели при минимальном числе обучаемых параметров (<0,2%). Работа вносит вклад в развитие NLP-инструментария для узбекского языка и может служить основой для дальнейших исследований в данной области.

Список литературы

[1] *Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems (NeurIPS), 33, 12449–12460.*

[2] *Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of ICML, 369–376.*

[3] *Heafield, K. (2011). KenLM: Faster and smaller language model queries. Proceedings of the Sixth Workshop on Statistical Machine Translation, 187–197.*

[4] *NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., … & Yankovskaya, E. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.*

[5] *Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286.*

[6] *Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. Proceedings of ICML, 1764–1772.*

[7] *Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. Proceedings of ICASSP, 4960–4964.*

[8] *Hasan, T., Bhattacharjee, A., Islam, Md. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., … & Shahriyar, R. (2021). XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. Findings of ACL-IJCNLP, 4693–4703.*

[9] *https://huggingface.co/datasets/yakhyo/mozilla-common-voice-uzbek*

[10] *Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. Proceedings of Interspeech, 2613–2617.*

[11] *Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Proceedings of ICLR.*

[12] *Kahn, J., Lee, A., & Hannun, A. (2022). pyctcdecode: A fast and flexible CTC decoder for speech recognition. GitHub. https://github.com/kensho-technologies/pyctcdecode*

[13] *Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. Advances in NeurIPS, 28, 1693–1701. [CNN/DailyMail dataset]*

[14] *Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. Proceedings of the ACL Workshop on Text Summarization Branches Out, 74–81.*

ҚМ АА Куәлік нөмірі: **KZ45VPY00102718** — ҚР Мәдениет және Ақпарат министрлігі