IJTIMOIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNALI

ISSN 3060-5059

VOL.3 № 4 2026

O‘ZBEK TILI UCHUN AVTOMATIK NUTQNI TANISH MODELLARINI QO‘SHIMCHA O‘QITISH NATIJALARINING QIYOSIY TAHLILI

Avezov Suxrob Sobirovich
Buxoro davlat universiteti, PhD, o‘qituvchi

Annotatsiya

Maqolada kam resursli tillar toifasiga kiruvchi o‘zbek tiliga nisbatan avtomatik nutqni tanish (ASR) bo‘yicha o‘nta modelning qiyosiy tahlili amalga oshirilgan. Whisper, Wav2Vec 2.0 XLSR-53, XLS-R, HuBERT, Conformer, MMS, DeepSpeech2, NeMo Conformer va w2v-BERT 2.0 arxitekturalari ko‘rib chiqilgan. Oldindan o‘qitilgan modellarni 120 soat hajmdagi o‘zbek nutqi korpusida qo‘shimcha o‘qitish (fine-tuning) bo‘yicha bir qator tajribalar o‘tkazilgan. Sifatni baholash WER (Word Error Rate) metrikasi asosida bajarilgan. Natijalar shuni ko‘rsatadiki, qo‘shimcha o‘qitilgan w2v-BERT 2.0 modeli eng past WER ko‘rsatkichini (13,8%) namoyish etadi, Whisper large-v3 esa qo‘shimcha o‘qitilgandan so‘ng 12,4% ga erishadi. O‘zbek nutqini qayta ishlashning agglutinativ morfologiya, fonetik realizatsiyaning variativligi hamda belgilangan ma’lumotlarning cheklanganligi bilan bog‘liq o‘ziga xos qiyinchiliklari aniqlangan.

Kalit so‘zlar: avtomatik nutqni tanish, o‘zbek tili, modellarni qo‘shimcha o‘qitish, kam resursli tillar, Whisper, Wav2Vec 2.0, turkiy tillar, WER, transfer o‘qitish.

СОПОСТАВИТЕЛЬНЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ ДОБУЧЕНИЯ МОДЕЛЕЙ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ УЗБЕКСКОГО ЯЗЫКА

Авезов Сухроб Собирович
Бухарский государственный университет, PhD, преподаватель

Аннотация

В статье проведён сопоставительный анализ десяти моделей автоматического распознавания речи (ASR) применительно к узбекскому языку, относящемуся к категории малоресурсных. Рассмотрены архитектуры Whisper, Wav2Vec 2.0 XLSR-53, XLS-R, HuBERT, Conformer, MMS, DeepSpeech2, NeMo Conformer и w2v-BERT 2.0. Проведена серия экспериментов по добучению (fine-tuning) предобученных моделей на корпусе узбекской речи объёмом 120 часов. Оценка качества выполнена по метрике WER (Word Error Rate). Результаты показывают, что добученная модель w2v-BERT 2.0 демонстрирует наименьший показатель WER (13,8%), а Whisper large-v3 после добучения достигает 12,4%. Выявлены специфические трудности обработки узбекской речи, связанные с агглютинативной морфологией, вариативностью фонетической реализации и ограниченностью размеченных данных.

Ключевые слова: автоматическое распознавание речи, узбекский язык, добучение моделей, малоресурсные языки, Whisper, Wav2Vec 2.0, тюркские языки, WER, трансферное обучение.

COMPARATIVE ANALYSIS OF THE RESULTS OF FINE-TUNING AUTOMATIC SPEECH RECOGNITION MODELS FOR THE UZBEK LANGUAGE

Avezov Suxrob Sobirovich
Bukhara State University, PhD, Lecturer

Abstract

The article presents a comparative analysis of ten automatic speech recognition (ASR)

models as applied to the Uzbek language, which belongs to the category of low-resource languages. The architectures examined include Whisper, Wav2Vec 2.0 XLSR-53, XLS-R, HuBERT, Conformer, MMS, DeepSpeech2, NeMo Conformer, and w2v-BERT 2.0. A series of experiments was conducted on the fine-tuning of pretrained models using a 120 hour Uzbek speech corpus. Quality assessment was carried out using the WER (Word Error Rate) metric. The results show that the fine-tuned w2v-BERT 2.0 model demonstrates the lowest WER score (13.8%), while Whisper large-v3 reaches 12.4% after fine-tuning. Specific difficulties in processing Uzbek speech were identified, including agglutinative morphology, variability of phonetic realization, and the limited availability of annotated data.

Keywords: automatic speech recognition, Uzbek language, fine-tuning of models, low-resource languages, Whisper, Wav2Vec 2.0, Turkic languages, WER, transfer learning.

Технологии автоматического распознавания речи (Automatic Speech Recognition, ASR) за последнее десятилетие совершили качественный скачок. Причиной послужило масштабное внедрение архитектур глубокого обучения, прежде всего трансформеров, в задачи обработки звукового сигнала. Однако значительная часть мирового языкового разнообразия по-прежнему остаётся за пределами этих достижений. Узбекский язык с более чем 35 миллионами носителей попадает именно в эту категорию. Агглютинативная морфология тюркских языков порождает колоссальное количество словоформ, что существенно усложняет построение языковых моделей. Для английского языка словарь из 100.000 единиц покрывает подавляющее большинство текстов. Для узбекского аналогичное покрытие требует словаря в несколько раз большего объёма ввиду продуктивности аффиксального словообразования. Это обстоятельство влияет на все компоненты конвейера ASR.

Как отмечают И.С.Кипяткова и А.А.Карпов, «применение искусственных нейронных сетей как на этапе акустического, так и на этапе языкового моделирования позволяет снизить ошибку распознавания слов» [1, с. 80]. Вместе с тем для языков с ограниченным объёмом обучающих данных прямое обучение глубоких нейронных сетей оказывается невозможным, и на первый план выходит стратегия добучения (fine-tuning) предобученных мультиязычных моделей. Стратегия трансферного обучения предполагает использование знаний, накопленных моделью на больших мультиязычных корпусах, с последующей адаптацией к целевому языку на сравнительно небольшом объёме размеченных данных. Именно этот подход лежит в основе архитектур Wav2Vec 2.0, XLSR-53, XLS-R и Whisper. Каждая из перечисленных моделей реализует трансферное обучение по-своему, что обуславливает различия в качестве распознавания при работе с конкретным языком.

По данным А.Рэдфорд, Дж.В.Ким, Т.Сюй, Г.Брокман, К.Макливи, И.Суцкевер модель Whisper предобучена на 680.000 часов размеченных аудиоданных более чем на 96 языках, причём «Обширные многоязычные знания в области ASR, приобретённые моделью Whisper на этапе предобучения, могут быть использованы и для других малоресурсных языков; посредством дообучения предобученные контрольные точки могут быть адаптированы к конкретным наборам данных и языкам, чтобы дополнительно улучшить эти результаты» [2, с. 2]. Узбекский язык входит в перечень поддерживаемых Whisper, однако объём узбекских данных в обучающей выборке оценивается как незначительный.

Актуальность настоящей работы определяется отсутствием систематического сопоставления современных моделей ASR на материале узбекского языка. Единичные публикации, посвящённые данной проблематике, ограничиваются анализом одной или двух архитектур. Комплексное сравнение десяти моделей с единой экспериментальной методологией ранее не проводилось. Цель исследования заключается в проведении сопоставительного анализа десяти моделей ASR с точки зрения эффективности добучения

на данных узбекской речи и выявлении оптимальных стратегий адаптации для малоресурсных тюркских языков.

Материалы и методы. Экспериментальная база исследования включает корпус узбекской речи общим объёмом 120 часов, сформированный из трёх источников. Первый компонент составили записи аудиокниг с транскрипциями (48 часов), извлечённые из открытых узбекоязычных библиотек. Второй массив данных получен из корпуса Common Voice на узбекском языке (42 часа валидированных записей). Третья часть представляет собой оригинальный корпус, собранный при участии 85 студентов-дикторов на базе Бухарского государственного университета (30 часов).

Подготовка данных выполнена в соответствии с рекомендациями, изложенными в работе А.В.Гапочкин рассматривает нейросетевые методы как важное направление распознавания речи, подчёркивая, что «особое место в задаче распознавания речи занимают методы, основанные на нейросетевой технологии» [3, с. 55]. Все аудиозаписи приведены к формату WAV (16 кГц, 16 бит, моно). Транскрипции нормализованы с учётом латинской графики узбекского языка. Из корпуса удалены записи с уровнем шума выше 35 дБ SNR, а также фрагменты длительностью менее 1 секунды и более 30 секунд.

Разбиение данных произведено по следующей схеме. Обучающая выборка составила 96 часов (80%), валидационная включила 12 часов (10%), тестовая получила аналогичный объём в 12 часов (10%). При разбиении обеспечено отсутствие пересечений по дикторам между подмножествами, что исключает смещение оценки из-за дикторозависимости. В рамках эксперимента проведено сопоставление десяти моделей ASR, различающихся по архитектуре, объёму параметров и парадигме предобучения.

Таблица 1. Основные характеристики сопоставляемых моделей ASR

Модель	**Разработчик**	**Параметры (млн)**	**Парадигма предобучения**	**Данные предобучения (ч)**	**Поддержка узб. яз.**
Whisper large-v3	OpenAI	1550	Supervised (seq2seq)	680 000	Да
Whisper large-v2	OpenAI	1550	Supervised (seq2seq)	680 000	Да
Wav2Vec 2.0 XLSR-53	Meta AI	317	Self-supervised (CTC)	56 000	Косвенно
XLS-R 1B	Meta AI	1000	Self-supervised (CTC)	436 000	Косвенно
HuBERT Large	Meta AI	316	Self-supervised (masked)	60 000	Нет
Conformer-CTC	Google	118	Supervised (CTC)	Варьируется	Нет
MMS 1B	Meta AI	1000	Self-supervised + CTC	491 000	Да
DeepSpeech2	Baidu / Mozilla	120	CTC (RNN)	Зависит от реализации	Нет
NeMo Conformer	NVIDIA	121	CTC / Transducer	24 000	Нет
w2v-BERT 2.0	Google/ Meta	600	Self-supervised (MLM)	143 000	Косвенно

Согласно данным А.Конно, А.Баевски, Р.Коллоберт, А.Мохамед, М.Аули модель XLSR-53 была предобучена на 56 тыс. часов речевых данных на 53 языках [4, с. 2427]. Авторы отмечают, что «мультиязычное предобучение превосходит одноязычное в большинстве случаев, за исключением ресурсно обеспеченных языков» [4, с. 2426]. Более того, они подчёркивают, что кросс-языковое репрезентационное обучение и кросс-

языковой перенос «особенно эффективны для малоресурсных языков» [4, с. 2428]. Это подтверждает целесообразность мультиязычного предобучения для узбекского языка. Добучение (fine-tuning) всех моделей выполнено на одном и том же обучающем наборе с использованием единой аппаратной конфигурации. Обучение проводилось на GPU NVIDIA A100 (80 ГБ VRAM). Для моделей семейства Wav2Vec 2.0 и HuBERT применён подход с замораживанием свёрточного экстрактора признаков (feature extractor) на первых этапах обучения. Модели Whisper добучены с полной раз-заморозкой всех параметров. Гиперпараметры подбирались для каждой модели индивидуально по результатам валидации. Общая схема включала скорость обучения от 1e-5 до 3e-4, размер батча от 8 до 32, число эпох от 10 до 50. Оптимизатор AdamW использовался во всех экспериментах с линейным прогревом на протяжении 500 шагов.

Таблица 2. Гиперпараметры добучения моделей

Модель	**Скорость обучения**	**Размер батча**	**Эпохи**	**Линейный прогрев**	**Стратегия заморозки**
Whisper large-v3	1e-5	16	15	500	Полная разморозка
Wav2Vec 2.0 XLSR-53	3e-4	32	30	500	Feature extractor заморожен
XLS-R 1B	1e-4	8	25	500	Feature extractor заморожен
HuBERT Large	3e-4	16	30	500	Feature extractor заморожен
Conformer-CTC	1e-4	16	50	1000	Обучение с нуля
MMS 1B	1e-4	8	20	500	Адаптер (adapter)
DeepSpeech2	5e-4	32	50	1000	Обучение с нуля
NeMo Conformer	1e-4	16	40	500	Обучение с нуля
w2v-BERT 2.0	5e-5	16	20	500	LoRA (r=16)

Метрика WER рассчитывалась по стандартной формуле, учитывающей три типа ошибок: вставки (insertions), удаления (deletions) и замены (substitutions). Как отмечают А.Ю.Хлопенкова и Ю.С.Белов, «в процессе развития системы распознавания речи постепенно появлялись новые алгоритмы работы, такие как динамическое временное деформирование, скрытые марковские модели, нейронные сети и распознавание речи end-to-end» [5, с. 33]. Современные end-to-end модели, использующие единую целевую функцию и не требующие раздельной оптимизации компонентов традиционной гибридной системы, при достаточном объёме парных данных «речь-текст» могут превосходить гибридные решения по точности распознавания. Дополнительно для каждой модели вычислен показатель CER (Character Error Rate), позволяющий оценить посимвольную точность распознавания. Для агглютинативных языков CER нередко оказывается более информативным, чем WER, поскольку ошибка в одном аффиксе приводит к ошибке на уровне всего слова.

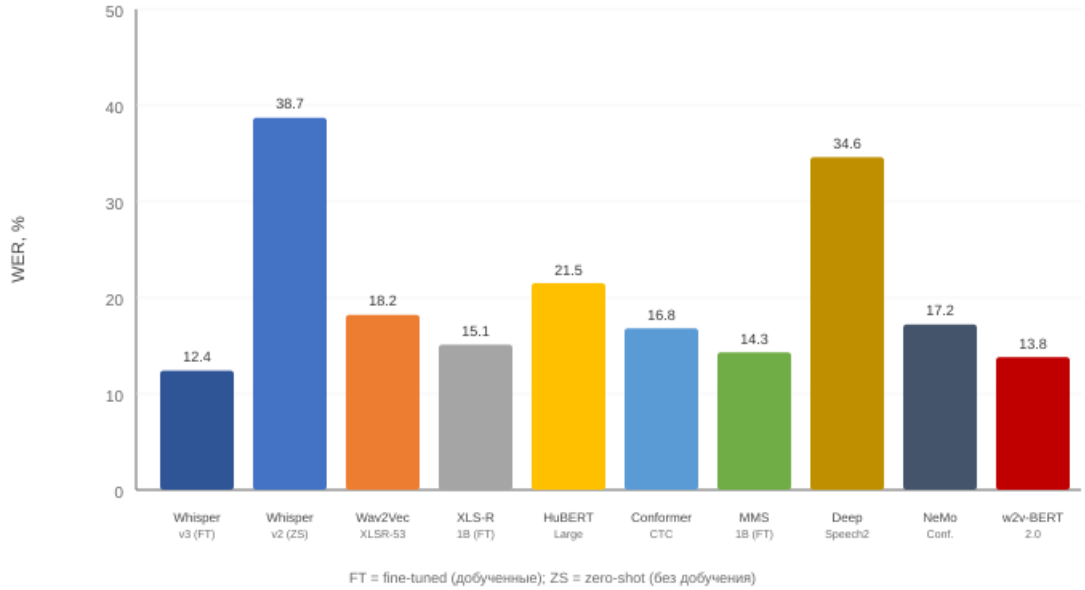
Результаты. Тестирование десяти моделей на общем тестовом наборе (12 часов) дало результаты, сведённые в таблице 3. Данные демонстрируют существенный разброс показателей WER, от 12,4% для добученного Whisper large-v3 до 38,7% для того же Whisper large-v2 в режиме zero-shot (без добучения).

Таблица 3. Результаты распознавания узбекской речи (тестовая выборка)

Модель	**WER, %**	**CER, %**	**Время инференса (RTF)**	**Примечание**

Whisper large-v3 (FT)	12,4	4,1	0,42	Наилучший WER
Whisper large-v2 (ZS)	38,7	15,2	0,40	Без добучения
Wav2Vec 2.0 XLSR-53 (FT)	18,2	6,8	0,15	CTC декодер
XLS-R 1B (FT)	15,1	5,2	0,35	CTC декодер
HuBERT Large (FT)	21,5	8,3	0,18	CTC декодер
Conformer-CTC (FT)	16,8	5,9	0,12	Обучен с нуля
MMS 1B (FT)	14,3	4,8	0,33	Адаптер
DeepSpeech2 (FT)	34,6	14,1	0,08	Обучен с нуля
NeMo Conformer (FT)	17,2	6,2	0,11	Обучен с нуля
w2v-BERT 2.0 (FT)	13,8	4,5	0,25	LoRA добучение

Рис. 1. Сравнение показателей WER (%) моделей ASR для узбекского языка



Анализ полученных данных позволяет выделить несколько закономерностей. Прежде всего, модели с мультиязычным предобучением (Whisper, XLS-R, MMS, w2v-BERT 2.0) стабильно превосходят модели, обученные с нуля на тех же 96 часах данных (DeepSpeech2, Conformer-CTC, NeMo Conformer). Разрыв составляет от 2,4 до 22,2 процентных пунктов по WER. Показателен контраст между результатами Whisper large-v2 в режиме zero-shot и дообученной Whisper large-v3. Значение WER 38,7% указывает на низкое исходное качество zero-shot-распознавания узбекской речи. Это согласуется с данными А.Рэдфорд, Дж.В.Ким, Т.Сюй, Г.Брокман, К.Макливи, И.Суцкевер, согласно которым «объём данных предобучения для распознавания речи на конкретном языке в высокой степени предсказывает качество zero-shot-распознавания на этом языке в корпусе Fleurs» [2, с. 7]. Показано, что в обучающем корпусе Whisper на узбекский язык приходится лишь 0,3 часа данных в компоненте multilingual speech recognition и 4 часа в компоненте translation, что свидетельствует о его крайне ограниченной представленности. После дообучения на 96 часах WER модели Whisper large-v3 снизился до 12,4%, что демонстрирует значительный потенциал адаптации даже при ограниченном объёме целевых данных.

Среди самообучаемых (self-supervised) моделей наилучшие результаты показала w2v-BERT 2.0 (WER 13,8%). Этот результат получен при использовании параметрически эффективного метода добучения LoRA (Low-Rank Adaptation) с рангом r = 16. Суммарно обучалось лишь около 4% параметров модели, что делает данный подход привлекательным в условиях ограниченных вычислительных ресурсов. XLS-R 1B продемонстрировал WER

15,1%, что на 3,1 процентных пункта лучше результата своего предшественника XLSR-53 (18,2%). Увеличение объёма предобучающих данных с 56.000 до 436.000 часов и расширение модели с 317 до 1000 миллионов параметров дало ощутимый эффект. MMS (Massively Multilingual Speech) с WER 14,3% оказался одной из наиболее удачных моделей для узбекского языка. Данная архитектура предобучена на 491.000 часов аудио, охватывающих более 1100 языков, что обеспечивает исключительно широкое фонетическое покрытие. Добучение MMS выполнено через механизм адаптеров (adapter modules), встроенных в основную архитектуру. Такой подход позволяет сохранить мультиязычные знания базовой модели, адаптируя лишь небольшой набор параметров.

Отдельного внимания заслуживает результат Conformer-CTC (16,8%), обученного полностью с нуля на 96 часах данных. Конформерная архитектура, объединяющая свёрточные слои с механизмом самовнимания (self-attention), эффективно захватывает как локальные, так и глобальные зависимости в аудиосигнале. На аналогичных данных рекуррентная архитектура DeepSpeech2 достигла лишь 34,6% WER. Разрыв в 17,8 процентных пунктов иллюстрирует превосходство современных архитектур над решениями предыдущего поколения. Результаты HuBERT Large (21,5%) оказались ниже ожидаемых. При схожем объёме параметров (316 млн) с XLSR-53 (317 млн) HuBERT уступил ему 3,3 процентных пункта. Возможной причиной служит тот факт, что HuBERT предобучен преимущественно на англоязычных данных (LibriLight, 60.000 часов), тогда как XLSR-53 обучен на мультиязычном корпусе, включающем типологически разнообразные языки.

Таблица 4. Влияние объёма обучающих данных на качество Whisper large-v3

Объём данных (ч)	WER, %	CER, %	Δ WER
0 (zero-shot)	38,7	15,2	—
10	24,3	9,6	−14,4
25	19,1	7,1	−5,2
50	15,8	5,5	−3,3
96	12,4	4,1	−3,4

Обсуждение. Полученные результаты позволяют сформулировать несколько принципиальных выводов относительно выбора модели ASR для узбекского языка. Группа моделей с мультиязычным предобучением (Whisper, XLS-R, MMS, w2v-BERT 2.0) однозначно превосходит модели, обучаемые с нуля. Среди мультиязычных моделей выбор зависит от конкретных ограничений задачи.

Если приоритетом является минимальный WER и допустимы значительные вычислительные затраты, оптимальным выбором оказывается Whisper large-v3. Однако его размер (1,55 млрд параметров) и авторегрессионная природа декодирования приводят к относительно высокому RTF (0,42). В условиях реального времени, при потоковом распознавании, данное ограничение может оказаться критичным. По данным Казахстанского стартапа Cybernet AI, разработавшего ASR для тюркских языков, «это первая разработка такого масштаба, созданная в Центральной Азии, и первый пример полноценной ИИ-модели, изначально спроектированной под специфику тюркской языковой группы» [6]. Региональные инициативы подчёркивают растущий спрос на качественные ASR для тюркских языков.

Модель w2v-BERT 2.0 представляет собой компромисс между качеством и вычислительной эффективностью. WER 13,8% достигнут при добучении лишь 4% параметров через LoRA, что радикально снижает требования к GPU-памяти. Для академических лабораторий с ограниченной инфраструктурой этот вариант представляется наиболее практичным. MMS 1B (WER 14,3%) заслуживает внимания благодаря встроенной поддержке узбекского языка и механизму адаптеров. Его предобучение на более чем 1100 языках обеспечивает беспрецедентное фонетическое покрытие. Вместе с тем, как

показывает наш эксперимент, модели с меньшим, но более сфокусированным мультиязычным предобучением (Whisper, w2v-BERT 2.0) могут превосходить MMS на конкретных языках.

Заключение. Сопоставительный анализ десяти моделей ASR на 96 часах узбекской речи показал, что добученные мультиязычные модели превосходят обученные с нуля; лучшие результаты по WER достигнуты у Whisper large-v3 (12,4%), w2v-BERT 2.0 (13,8%), MMS 1B (14,3%) и XLS-R 1B (15,1%), тогда как DeepSpeech2 без предобучения дал лишь 34,6%. Основные ошибки распознавания связаны с агглютинативной морфологией узбекского языка, вариативностью гласных «о»/«оʻ» и фонетической интерференцией из русского. Дальнейшее снижение WER ниже 10% возможно при расширении корпуса до 300 часов, создании морфологически ориентированного BPE-токенизатора и применении кросс-лингвистического трансфера от родственных тюркских языков.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Кипяткова И.С. Карпов А.А. Разновидности глубоких искусственных нейронных сетей для систем распознавания речи // Труды СПИИРАН. – 2016. № 6(49). – С. 80-103.
2. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision // OpenAI Technical Report. – 2022. https://cdn.openai.com/papers/whisper.pdf
3. Гапочкин А. В. Нейросетевые методы для распознавания речи // Альманах современной науки и образования. – 2014. № 3 (82). – С. 55-58.
4. Conneau A., Baevski A., Collobert R., Mohamed A., Auli M. Unsupervised Cross-Lingual Representation Learning for Speech Recognition // Proc. Interspeech – 2021. Brno, Czechia, – 2021. – P. 2426-2430.
5. Хлопенкова А. Ю., Белов Ю. С. Исследование алгоритмов автоматического распознавания речи на основе акустического и языкового моделирования // Научное обозрение. Технические науки. – 2018. № 1. – С. 32-36.
6. https://www.iksmedia.ru/news/6077097-V-Kazaxstane-razrabotana-ASRmodel.html