ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРИНЯТИЕ РЕШЕНИЙ
ARTIFICIAL INTELLIGENCE AND DECISION MAKING

УДК 004.8 **DOI: 10.18413/2518-1092-2026-11-1-0-3**

Тимченко О.В.[1]
Алексеева Д.К.[2]
Абрегова З.Х.[1]
Гречко В.А.[1]

АРХИТЕКТУРА СИСТЕМЫ ASR ДЛЯ АГГЛЮТИНАТИВНЫХ ЯЗЫКОВ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

[1)] Пятигорский государственный университет,
пр-т Калинина, 9, г. Пятигорск, Ставропольский край, 357532, Россия
[2)] Университет ИТМО, пр-т Кронверкский, 49, лит. А, г. Санкт-Петербург, 197101, Россия

e-mail: gorbachenkotim@mail.ru, dashaalekseeva08.20@gmail.com, zalinaabregova@mail.ru, Lera197689@yandex.ru

Аннотация
Актуальность темы исследования обусловлена необходимостью преодоления цифрового неравенства, которое особенно остро проявляется в отношении малоресурсных языков. В то время как носители распространенных языков активно пользуются голосовыми помощниками, системами транскрибации и другими речевыми технологиями, малочисленные коренные народы остаются за бортом цифрового прогресса. Это неравенство лишает людей доступа к современным средствам коммуникации, образования и информации на родном языке, что ведет к их дальнейшей маргинализации и ускоряет процесс языкового вымирания. Разработка специализированных решений для автоматического распознавания речи (ASR) в условиях ограниченных данных является ключевым шагом на пути расширения технологической доступности. В статье рассматривается проблема разработки систем автоматического распознавания речи (ASR) для малоресурсных языков, в частности, кабардинского. Представлен комплексный подход, включающий адаптацию массово многоязычной модели MMS (Massively Multilingual Speech), предобработку данных, а также разработку и интеграцию языковых моделей для постобработки. Основное внимание уделено архитектуре модели MMS, основанной на Wav2Vec 2.0, и её модификации с использованием языково-специфических адаптеров (LSAH), что позволяет эффективно дообучать модель на ограниченных наборах данных. Описаны этапы предобработки аудио и текстовых данных. Рассмотрены архитектуры и результаты применения n-граммных (3-gram, 5-gram) и нейросетевой (mT5-base) языковых моделей для коррекции ошибок в выводе ASR. Практическая значимость работы подтверждена созданием рабочей open-source системы с веб-интерфейсом на платформе Hugging Face Spaces, демонстрирующей возможность построения эффективных ASR-решений для миноритарных языков.

Ключевые слова: автоматическое распознавание речи (ASR); малоресурсные языки; кабардинский язык; MMS (Massively Multilingual Speech); Wav2Vec 2.0; адаптеры; языковые модели; постобработка; n-граммы; mT5

Для цитирования: Тимченко О.В., Алексеева Д.К., Абрегова З.Х., Гречко В.А. Архитектура системы ASR для агглютинативных языков с ограниченными ресурсами // Научный результат. Информационные технологии. – Т.11, №1, 2026. – С. 20-28. DOI: 10.18413/2518-1092-2026-11-1-0-3

Timchenko O.V.[1]
Alekseeva D.K.[2]
Abregova Z.Kh.[1]
Grechko V.A.[1]

SYSTEM ARCHITECTURE FOR ASR OF AGGLUTINATIVE LOW-RESOURCE LANGUAGES

[1)] Pyatigorsk State University, 9 Kalinin Ave, Pyatigorsk, Stavropol Krai, 357532, Russia
[2)] ITMO University, 49 Letter A Kronverksky Ave, St. Petersburg, 197101, Russia

e-mail: gorbachenkotim@mail.ru, dashaalekseeva08.20@gmail.com, zalinaabregova@mail.ru, Lera197689@yandex.ru

Abstract
The relevance of the research is driven by the need to overcome the digital divide, which is particularly acute for low-resource languages. While speakers of widely spoken languages actively use voice assistants, transcription systems, and other speech technologies, small indigenous peoples are left behind in the digital progress. This inequality deprives people of access to modern means of communication, education, and information in their native language, leading to their further marginalization and accelerating the process of language extinction. The development of specialized solutions for automatic speech recognition (ASR) under low-resource conditions is a key step towards expanding technological accessibility. The article addresses the problem of developing automatic speech recognition (ASR) systems for low-resource languages, specifically Kabardian. It presents a comprehensive approach, including the adaptation of the Massively Multilingual Speech (MMS) model, data preprocessing, as well as the development and integration of language models for post-processing. The main focus is on the MMS model architecture, based on Wav2Vec 2.0, and its modification using Language-Specific Adapter Heads (LSAH), which enables efficient fine-tuning of the model on limited datasets. The stages of audio and text data preprocessing are described. The architectures and results of applying n-gram (3-gram, 5-gram) and neural network (mT5-base) language models for correcting errors in the ASR output are considered. The practical significance of the work is confirmed by the creation of a functional open-source system with a web interface on the Hugging Face Spaces platform, demonstrating the feasibility of building effective ASR solutions for minority languages.
Keywords: automatic speech recognition (ASR); low-resource languages; Kabardian language; MMS (Massively Multilingual Speech); Wav2Vec 2.0; adapters; language models; post-processing; n-grams; mT5
For citation: Timchenko O.V., Alekseeva D.K., Abregova Z.Kh., Grechko V.A. System Architecture for ASR of Agglutinative Low-Resource Languages // Research result. Information technologies. – T.11, №1, 2026. – P. 20-28. DOI: 10.18413/2518-1092-2026-11-1-0-3

ВВЕДЕНИЕ

Широкое распространение технологий голосового ввода, таких как голосовые помощники, системы транскрибации и средства коммуникации, обуславливает растущую актуальность задач автоматического распознавания речи (Automatic Speech Recognition, ASR). Однако подавляющее большинство существующих ASR-систем ориентировано на многоресурсные языки (английский, китайский, испанский и др.), оставляя без поддержки около 7000 языков коренных народов, которые относятся к категории малоресурсных. Отсутствие крупных размеченных аудиотекстовых корпусов, слабая коммерческая привлекательность и высокая стоимость создания датасетов являются основными препятствиями для разработки ASR для таких языков, что усугубляет проблему цифрового неравенства [3, 8, 13].

Кабардино-черкесский язык (далее кабардинский), относящийся к абхазо-адыгской семье, является ярким примером малоресурсного языка со сложной агглютинативной морфологией и эргативным строем. Эти лингвистические особенности создают дополнительные проблемы для ASR-систем, приводя к ошибкам морфологического характера и сегментации слов.

Целью исследования является проектирование интеллектуальной системы ASR для кабардинского языка, включающей дообучение массово многоязычной модели MMS и создание

языковых моделей для постобработки её вывода. Основное внимание в статье уделяется архитектурным аспектам предлагаемого решения, в частности, детальному рассмотрению модели MMS и её адаптации под целевой язык.

ОСНОВНАЯ ЧАСТЬ

Современные ASR-системы эволюционировали от гибридных моделей (GMM-HMM, DNN-HMM) к сквозным (end-to-end) архитектурам, таким как модели на основе CTC (Connectionist Temporal Classification) [4] и трансформеров. End-to-end подход, при котором модель обучается как единое целое для прямого преобразования аудио в текст, показал высокую эффективность, особенно в сочетании с механизмами внимания (attention).

Для малоресурсных языков наиболее перспективным является использование предобученных многоязычных моделей, способных к переносу знаний (cross-lingual transfer) [12, 14]. Сравнительный анализ существующих моделей, таких как XLS-R, HuBERT, mHuBERT-147 и Meta-Whisper, показал, что модель MMS (Massively Multilingual Speech) обладает рядом преимуществ для адаптации под языки с ограниченными ресурсами [11].

MMS предобучена на колоссальном корпусе объемом 491 тысяча часов неразмеченной речи на 1406 языках, что является наибольшим охватом среди рассмотренных аналогов [11]. Это обеспечивает модели богатые начальные представления о разнообразии речевых паттернов. Ключевой особенностью MMS, определившей её выбор для данной работы, является адаптерная архитектура (LSAH – Language-Specific Adapters, Head and Fine-Tuning), позволяющая эффективно дообучать модель на новых языках, замораживая основную часть параметров и обучая лишь небольшое количество языково-специфичных слоев. Это значительно снижает вычислительные затраты и риск переобучения на малых данных.

Архитектура MMS базируется на модели Wav2Vec 2.0 [5], которая состоит из нескольких ключевых компонентов:

1. Сверточный энкодер (CNN Encoder) принимает на вход сырой аудиосигнал (волновую форму) и преобразует его в последовательность латентных речевых представлений. Эти представления захватывают локальные акустические особенности.

2. Трансформерный контекстуализатор (Transformer Context Network) принимает латентные представления и обогащает их контекстной информацией со всей последовательности с помощью механизма самовнимания (self-attention). На выходе получаются контекстуализированные представления.

3. В процессе предобучения часть латентных представлений маскируется. Модель обучается предсказывать квантованные (дискретные) представления маскированных участков на основе контекстуализированных представлений от трансформера. Квантование производится с помощью модуля квантования (Quantization Module), который проецирует непрерывные векторы на дискретный набор прототипов (codebook). Это позволяет модели изучать дискретные речевые единицы без явной разметки.

4. Обучение осуществляется с использованием контрастивной функции потерь (Contrastive Loss). Модель учится отличать квантованное представление маскированного временного шага (положительный пример) от дистракторов (отрицательных примеров) – квантованных представлений из других частей аудио.

Эта архитектура позволяет Wav2Vec 2.0 извлекать обобщенные речевые представления из неразмеченных данных.

MMS расширяет подход Wav2Vec 2.0 за счет обучения на данных с 1406 языков. Для эффективной адаптации этой огромной модели к конкретным задачам (ASR, идентификация языка, синтез речи) и языкам используется механизм адаптеров [9].

Адаптеры – это небольшие, встраиваемые нейросетевые модули, которые добавляются в каждый блок трансформера предобученной модели MMS. Конкретно, они размещаются после feed-forward слоя внутри трансформерного блока.

Структура одного адаптера включает:

1. LayerNorm (Нормализация слоя) нормализует входной вектор от feed-forward слоя.
2. Down-Projection (Понижение размерности) – линейный слой, который проецирует входной вектор высокой размерности (например, 1024) в пространство значительно меньшей размерности (например, 256). Это необходимо для сокращения числа параметров.
3. ReLU (Rectified Linear Unit) – нелинейная функция активации.
4. Up-Projection (Восстановление размерности) – линейный слой, который проецирует вектор обратно в исходную высокую размерность (например, 1024).
5. Остаточное соединение (Residual Connection) используется на выходе адаптера, который суммируется с его исходным входом, что позволяет сохранить информацию, полученную на предыдущих этапах, и стабилизировать обучение.

Таким образом, адаптер учится преобразовывать признаки, специфичные для данного языка или задачи, не изменяя основные, общие для многих языков веса исходной модели.

Для адаптации MMS к задаче распознавания кабардинской речи был применен подход LSAH, который состоит из двух этапов:

1. Добавление и инициализация адаптеров. В предобученную модель MMS добавляются языково-специфические адаптеры для кабардинского языка. Основные параметры модели (веса CNN-энкодера и трансформерных блоков) замораживаются и не обновляются в процессе дообучения.

2. Тонкая настройка (Fine-Tuning) адаптеров и выходного слоя. В этом случае добавляется новый выходной слой (Linear Head), случайно инициализированный линейный классификатор, который отображает выходные представления трансформера в вероятности символов словаря кабардинского языка. Дообучению подвергаются только параметры адаптеров и этого выходного слоя. Общее количество обучаемых параметров составляет около 2 миллионов на язык, что примерно на 2% превышает размер исходной модели, делая процесс чрезвычайно эффективным с вычислительной точки зрения. Для обучения используется функция потерь CTC (Connectionist Temporal Classification), которая подходит для задач выравнивания последовательностей разной длины (как аудио, так и текстовых) без необходимости пошаговой разметки. Используется алгоритм Adam (Adaptive Moment Estimation), который сочетает в себе преимущества двух других методов – RMSProp и Momentum. Adam адаптивно вычисляет индивидуальные скорости обучения для каждого параметра модели, используя скользящие средние как, так и вторых моментов. Это позволяет алгоритму эффективно работать с разреженными градиентами, характерными для задач NLP и ASR, и обеспечивать стабильную и быструю сходимость даже на зашумленных данных, что критически важно при работе с ограниченными наборами данных для малоресурсных языков [7]. Данный подход позволяет достичь высокого качества распознавания, используя для дообучения менее 100 часов размеченных аудиоданных целевого языка.

Для дообучения модели были использованы три датасета кабардинской речи, представленные в таблице 1. Для всех датасетов предусмотрено разделение на обучающую и тестовую выборки. Данные kbd_speech и sixuxar_yijiri_mak7 разделены на обучающие и тестовые наборы (train/test), вручную размеченный корпус дополнительно использовался для валидации модели и оценки качества на длинных последовательностях.

Таблица 1

Характеристика датасетов

Table 1

Dataset Characteristics

Название	Источ-ник	Кол-во файлов	Длитель-ность аудио	Тип данных	Мета-данные	Формат аннотации
kbd_speech	HuggingFace	20555	2–3 сек/файл	Изолированные слова	Пол, страна, ID спикера	audio, transcription, gender, country, speaker_id
sixuxar_yijiri_mak7	HuggingFace	6579	5–15 сек/файл	Короткие фразы и предложения	–	audio, text
Собственный датасет	Собран вручную	55	20–90 сек/файл	Несколько предложений, абзацы	–	audio, text

В ходе исследования рассматривался вопрос о применении аугментации аудиоданных для повышения устойчивости модели. Однако было принято решение отказаться от её использования на данном этапе. Основной причиной послужил характер исходных данных. Все использованные датасеты (kbd_speech, sixuxar_yijiri_mak7, собственный датасет) содержали записи приемлемого акустического качества без фоновых шумов и артефактов. Поскольку целевой сценарий применения системы на начальном этапе также предполагает работу с «чистым» аудиовходом, аугментация, имитирующая шумы и искажения, могла бы необоснованно усложнить модель и сместить её внимание с фундаментальных для кабардинского языка паттернов – сложной фонетики и агглютинативной морфологии. Общее разнообразие данных (от слов к связной речи) способствовало более устойчивому обучению.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Следует отметить, что на момент проведения данного исследования отсутствовали какие-либо публично доступные или описанные в литературе системы автоматического распознавания речи кабардино-черкесского языка. Кабардинский язык, являясь классическим примером малоресурсного языка, не был представлен в финальных слоях классификации крупных многоязычных моделей, таких как XLS-R или Whisper, что делало их прямое применение без дообучения невозможным [16]. Предварительная оценка показала, что качество распознавания исходной моделью MMS до адаптации было неприемлемо низким (WER > 0.95, Word Error Rate). В связи с этим, ключевой задачей исследования создание и оценка работоспособной системы автоматического распознавания речи для кабардино-черкесского языка.

В результате проведены два эксперимента по токенизации.

Эксперимент 1 включал замену сложных диграфов и триграфов кабардинского языка на уникальные символы с целью уменьшения размера словаря. Итоговый словарь – 34 токена. Для коррекции ошибок в выводе ASR были разработаны 3-граммная и 5-граммная языковые модели, основанные на марковском предположении о зависимости вероятности слова от предыдущих n-1 слов. Анализ показал, что 3-граммная модель демонстрирует стабильную работу на всех датасетах. На kbd_speech наблюдалась низкая перплексия (873.57 с учетом количества неизвестных слов, Out-of-Vocabulary, OOV), что свидетельствует о хорошей адаптации к изолированным словам. На датасете sixuxar_yijiri_mak7 модель показала перплексию 89.68, эффективно обрабатывая короткие фразы. На собственном датасете с протяженными текстами перплексия возросла до 29521.04, что объясняется сложностью контекстного моделирования. 5-граммная модель проявила высокую

чувствительность к неизвестной лексике, особенно на kbd_speech. Увеличение порядка n-граммы не привело к улучшению качества, а напротив, усилило зависимость от объема обучающих данных.

В эксперименте 2 в качестве альтернативы n-граммным моделям была дообучена многоязычная модель mT5-base на синтетическом датасете из 220 тысяч пар «ошибочное предложение – эталон». Для генерации ошибок использовался специальный алгоритм аугментации, имитирующий характерные для ASR искажения, такие как фонетические замены, вставки, удаления символов и ошибки сегментации. Эксперимент основывался на гипотезе о целесообразности сохранения естественной орфографии кабардинского языка. Методология предобработки предусматривала минимальное вмешательство в исходные данные, в том числе очистку текстов от пунктуации, приведение к нижнему регистру, унификацию символа «I» и замену пробела на «*» без модификации диграфов и триграфов. Итоговый словарь включал 37 токенов, полностью соответствующих оригинальному кабардинскому алфавиту с сохранением всех многосимвольных графем.

Результаты показали, что стратегия без замен (Эксперимент 2) приводит к лучшим результатам WER = 0.4186, CER = 0.0912 (Character Error Rate), против WER = 0.8683, CER = 0.3089 после обратной замены в первом эксперименте. Это подтвердило важность сохранения оригинальных орфографических паттернов для агглютинативных языков.

Аудиофайлы были ресемплированы до 16 кГц, нормализованы и преобразованы в формат, пригодный для обработки моделью, с использованием процессора Wav2Vec2Processor из библиотеки Hugging Face Transformers.

Анализ вывода дообученной MMS показал, что основные ошибки связаны не с акустическим распознаванием, а с грамматическими искажениями и ошибками сегментации (слияние и разделение слов). Для их исправления были исследованы два типа языковых моделей (LM).

Были построены 3-граммная и 5-граммная LM с использованием библиотеки KenLM. Модели оценивались по перплексии и количеству неизвестных слов (OOV), результаты сравнения представлены в таблице 2.

Таблица 2

Сравнение n-gramm моделей

Table 2

Comparison of n-gram models

Датасет	3-gram LM (PPL incl. OOVs)	3-gram LM (PPL excl. OOVs)	3-gram LM (OOVs)	5-gram LM (PPL incl. OOVs)	5-gram LM (PPL excl. OOVs)	5-gram LM (OOVs)
kbd_speech	873.57	873.37	48	2164.92	49.3	15579
sixuxar_yijiri_mak7	89.68	69.22	1462	28145.41	1732.91	22343
Собственный	29521.04	8755.74	326	98517.99	5388.8	748

3-граммная модель показала значительно лучшую устойчивость и обобщающую способность на всех тестовых датасетах по сравнению с 5-граммной. Низкая перплексия на датасетах с короткими фразами (kbd_speech, sixuxar_yijiri_mak7) свидетельствовала о её адекватности для данного типа данных.

3-граммная модель была интегрирована в CTC-декодер с помощью библиотеки pyctcdecode для повторного ранжирования гипотез на этапе инференса. Это привело к существенному снижению итогового WER (до ≈0.318).

В качестве альтернативы была дообучена многоязычная модель mT5-base (300M параметров) для задачи исправления ошибок. Для обучения был синтезирован датасет из 220 тысяч пар «предложение с ошибкой – эталон». Ошибки генерировались искусственно (вставки, удаления, замены символов, ошибки сегментации).

На валидационной выборке модель показала высокое качество, подтвержденное показателями метрик Loss = 0.31, WER = 0.249, CER = 0.047. Однако при применении к реальным гипотезам MMS качество ухудшилось, что вызвало увеличение WER с 0.296 до 0.403. Это указывает на сдвиг домена

между синтетическими ошибками и реальными ошибками ASR, а также на возможную необходимость более длительного обучения.

Стабильность метрик оценивалась по согласованности показателей WER и CER на трех различных по характеру датасетах. Так, на датасете `kbd_speech` (изолированные слова) модель показала низкий CER (0.0763), на `sixuxar_yijiri_mak7` (короткие фразы) метрики оставались стабильными, а на наиболее сложном собственном датасете (связная речь) наблюдалось ожидаемое увеличение WER. Эта согласованность на разнородных данных указывает на то, что модель не подстроилась под специфику одного набора.

Риск переобучения был минимизирован за счет двух ключевых факторов:

1. Использование адаптерной архитектуры (LSAH)**.** При дообучении модели MMS были заморожены основные параметры (≈98% весов), что кардинально снизило количество обучаемых параметров (около 2 млн.) и, как следствие, риск запоминания данных.

2. Мониторинг процесса обучения. Динамика функции потерь (*train/loss*) демонстрировала плавное и устойчивое снижение без признаков расхождения с валидационной кривой.

Полученные результаты, в особенности значительное улучшение качества после интеграции 3-граммной LM, которое наблюдалось на всех тестовых выборках, позволяют считать выводы работы устойчивыми и воспроизводимыми в рамках использованного экспериментального протокола.

Алгоритм синтетической аугментации генерировал в основном локальные вставки, удаления и замены символов. Синтетические ошибки были по своей сути контекстно-независимыми. В реальных же транскрипциях многие ошибки распознавания были условными и зависели от окружающего контекста (например, фонетического сходства слов в предложении).

Модель mT5 на наш взгляд является избыточной для задачи локальной коррекции, которая, по сути, является редактированием. Склонность модели «переписывать» предложения, а не точечно исправлять ошибки, в условиях сдвига домена приводила к внесению новых, несвойственных исходной ASR-системе искажений.

Таким образом, основной причиной ухудшения распознавания можно рассматривать недостаточную репрезентативность синтетического датасета, не способного уловить системные и контекстно-обусловленные ошибки реальной ASR-системы. Это указывает на то, что для успешного применения нейросетевых LM для постобработки в будущем необходим сбор датасета.

Несмотря на это, подход демонстрирует потенциал, но требует дальнейшей работы со сбором датасета реальных ошибок.

Для демонстрации работы системы было развернуто веб-приложение на платформе Hugging Face Spaces с использованием фреймворка Gradio. Приложение позволяет пользователям загружать аудиофайлы (WAV/MP3) или записывать речь непосредственно в браузере и получать текстовую транскрипцию, полученную с помощью дообученной модели MMS с интегрированной 3-граммной языковой моделью.

Архитектура пайплайна инференса включает:

1. Загрузку аудио и предобработку (librosa).

2. Инференс модели MMS для получения логитов.

3.Декодирование логитов с использованием CTC-декодера, усиленного 3-граммной LM (pyctcdecode).

4. Вывод итоговой транскрипции пользователю.

ЗАКЛЮЧЕНИЕ

В работе представлена комплексная методика разработки системы ASR для малоресурсного кабардинского языка. Ключевым элементом системы является многоязычная модель MMS, архитектура которой на базе Wav2Vec 2.0 с языково-специфическими адаптерами доказала свою эффективность в условиях ограниченных данных. Адаптерный подход LSAH позволил провести успешное дообучение, минимизируя вычислительные затраты.

Эксперименты подтвердили, что для агглютинативных языков предпочтительнее стратегия токенизации с сохранением оригинальной орфографии. Показано, что даже простая 3-граммная языковая модель, интегрированная в декодер, значительно улучшает качество транскрипции за счет коррекции ошибок сегментации. Нейросетевая модель mT5 требует дополнительной доработки для работы с реальными ошибками ASR.

Практическим результатом работы является функционирующая open-source система с публичным веб-интерфейсом, что снижает барьер для цифровизации кабардинского языка и служит основой для будущих исследований в области ASR для других малоресурсных языков.

Список литературы

1. Алексеева Д.К. Технологии автоматического распознавания речи на малоресурсных миноритарных языках Северного Кавказа / Д.К. Алексеева, О.В. Тимченко // Наукосфера. – 2024.

2. Кипяткова И.С., Кагиров, И.А. Система автоматического распознавания карельской речи / И.С. Кипяткова, И.А. Кагирова // Информационно-управляющие системы. – 2023. – Т. 3. – С. 16-25.

3. Кузьмин Е.И. Современные проблемы сохранения и развития миноритарных языков в условиях многоязычия в России и в мире: пути решения и перспективы / Е.И. Кузьмин // Университетская книга. – 2022. – URL: https://www.unkniga.ru/kultura/13442-sovremennye-problemysohraneniya-i-razvitiyaminoritarnyh-yazykov-v-usloviyahmnogoyazychiya.html.

4. Orken M. Study of transformer-based end-to-end speech recognition system for Kazakh language / M. Orken, O. Dina, A. Keylan [et al.] // Sci Rep. – 2022. – Vol. 12. – Pp. 8337.

5. Baevski A. Wav2vec 2.0: a framework for self-supervised learning of speech representations / A. Baevski [et al.] // Advances in Neural Information Processing Systems. – 2020.

6. Boosting Wav2Vec2 with N-Grams in Transformers // Hugging Face Blog. – URL: https://huggingface.co/blog/wav2vec2with-ngram (дата обращения: 17.06.2025).

7. Wang H. Understanding knowledge transferability for transfer learning: a survey / H. Wang [et al.] // ACM Comput. Surv. – 2025. – Vol. 1, No. 1. – July. – 35 p. DOI: 10.1145/XXXXXXX.XXXXXXX.

8. Dialectal diversity and its effect on the language model landscape // Appen blog. – URL: https://www.appen.com/blog/pulseoflanguageevolution (дата обращения: 10.03.2025).

9. Fine-tuning MMS adapter models for multi-lingual ASR // Hugging Face Blog. – URL: https://huggingface.co/blog/mmsadapters. (дата обращения: 08.04.2025).

10. Hou W. Exploiting adapters for cross-lingual low-resource speech recognition / W. Hou [et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021.

11. Pratap V. Scaling speech technology to 1,000+ languages / V. Pratap [et al.] // Journal of Machine Learning Research. – 2024.

12. Protasov V. Super donors and super recipients: studying cross-lingual transfer between high-resource and low-resource languages / V. Protasov [et al.] // Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages. – 2024. – Pp. 94-108.

13. Supriyono A. Advancements in Natural Language Processing: Implications, Challenges, and Future Directions / A. Supriyono [et al.] // Telematics and Informatics Reports. – 2024. – Vol. 16. – Art. no. 100173. DOI: 10.1016/j.teler.

14. Transfer learning with Keras // Neurohive.io. – URL: https://neurohive.io/ru/tutorial/transfer-learningkeras (дата обращения: 23.02.2025).

15. Latif S. Transformers in Speech Processing: A Survey / S. Latif [et al.] // 2023. – URL: https://arxiv.org/abs/2303.11607.

16. Babu A. XLS-R: SelfSupervised Cross-Lingual Speech Representation Learning at Scale / A. Babu [et al.] // Proc. Interspeech. – 2022. – Pp. 2278-2282.

17. Xue L. MT5: A Massively Multilingual PreTrained TexttoText Transformer / L. Xue [et al.] // ArXiv preprint arXiv:2010.11934. – 2020.

References

1. Alekseeva D.K. Technologies of automatic speech recognition in low-resource minority languages of the North Caucasus / D.K. Alekseeva, O.V. Timchenko // Naukosphere. – 2024.

2. Kipyatkova I.S., Kagirov, I.A. System of automatic recognition of Karelian speech / I.S. Kipyatkova, I.A. Kagirova // Information and control systems. – 2023. – Vol. 3. – P. 16-25.

3. Kuzmin E.I. Modern problems of preservation and development of minority languages in the context of multilingualism in Russia and in the world: solutions and prospects / E.I. Kuzmin // University book. – 2022. – URL: https://www.unkniga.ru/kultura/13442-sovremennye-problemysohraneniya-i-razvitiyaminoritarnyh-yazykov-v-usloviyahmnogoyazychiya.html.

4. Orken M. Study of transformer-based end-to-end speech recognition system for Kazakh language / M. Orken, O. Dina, A. Keylan [et al.] // Sci Rep. – 2022. – Vol. 12. – Pp. 8337.

5. Baevski A. Wav2vec 2.0: a framework for self-supervised learning of speech representations / A. Baevski [et al.] // Advances in Neural Information Processing Systems. – 2020.

6. Boosting Wav2Vec2 with N-Grams in Transformers // Hugging Face Blog. – URL: https://huggingface.co/blog/wav2vec2with-ngram (date of access: 17.06.2025).

7. Wang H. Understanding knowledge transferability for transfer learning: a survey / H. Wang [et al.] // ACM Comput. Surv. – 2025. – Vol. 1, No. 1. – July. – 35 p. DOI: 10.1145/XXXXXXX.XXXXXXX.

8. Dialectal diversity and its effect on the language model landscape // Appen blog. – URL: https://www.appen.com/blog/pulseoflanguageevolution (date of access: 10.03.2025).

9. Fine-tuning MMS adapter models for multi-lingual ASR // Hugging Face Blog. – URL: https://huggingface.co/blog/mmsadapters. (date of access: 04/08/2025).

10. Hou W. Exploiting adapters for cross-lingual low-resource speech recognition / W. Hou [et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021.

11. Pratap V. Scaling speech technology to 1,000+ languages / V. Pratap [et al.] // Journal of Machine Learning Research. – 2024.

12. Protasov V. Super donors and super recipients: studying cross-lingual transfer between high-resource and low-resource languages / V. Protasov [et al.] // Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages. – 2024. – Pp. 94-108.

13. Supriyono A. Advancements in Natural Language Processing: Implications, Challenges, and Future Directions / A. Supriyono [et al.] // Telematics and Informatics Reports. – 2024. – Vol. 16. – Art. no. 100173. DOI: 10.1016/j.teler.

14. Transfer learning with Keras // Neurohive.io. – Access mode: https://neurohive.io/ru/tutorial/transfer-learningkeras (date of access: 23.02.2025).

15. Latif S. Transformers in Speech Processing: A Survey / S. Latif [et al.] // 2023. – URL: https://arxiv.org/abs/2303.11607.

16. Babu A. XLS-R: SelfSupervised Cross-Lingual Speech Representation Learning at Scale / A. Babu [et al.] // Proc. Interspeech. – 2022. – Pp. 2278-2282.

17. Xue L. MT5: A Massively Multilingual PreTrained TexttoText Transformer / L. Xue [et al.] // ArXiv preprint arXiv:2010.11934. – 2020.

Тимченко Ольга Викторовна, кандидат экономических наук, доцент, Пятигорский государственный университет, г. Пятигорск, Россия
Алексеева Дарья Константиновна, магистрант, Университет ИТМО, г. Санкт-Петербург, Россия
Абрегова Залина Хамидбиевна, ассистент, Пятигорский государственный университет, г. Пятигорск, Россия
Гречко Валерия Андреевна, ассистент, Пятигорский государственный университет, г. Пятигорск, Россия

Timchenko Olga Viktorovna, Candidate of Economic Sciences, Associate Professor, Pyatigorsk State University, Pyatigorsk, Russia
Alekseeva Darya Konstantinovna, Master's student, ITMO University, St. Petersburg, Russia
Abregova Zalina Khamidbievna, Assistant Professor, Pyatigorsk State University, Pyatigorsk, Russia
Grechko Valeriya Andreevna, Assistant Professor, Pyatigorsk State University, Pyatigorsk, Russia