СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ WAVE2VEC И ТРАДИЦИОННЫХ МЕТОДОВ HMM ДЛЯ РАСПОЗНАВАНИЯ КАРАКАЛПАКСКОЙ РЕЧИ

Н. Утеулиев
Д.ф.м.н, проф. Нукусский Государственный Технический Университет
Ж.Кудайбергенов
Докторант. Нукусский Государственный Технический Университет
Т. Кудайбергенов
Магистр. Нукусский Государственный Технический Университет

Аннотация. В данной статье проводится сравнительный анализ двух принципиально различных подходов к автоматическому распознаванию речи (Automatic Speech Recognition, ASR) применительно к низкоресурсному каракалпакскому языку: традиционного подхода на основе скрытых марковских моделей (Hidden Markov Models, HMM) и современного подхода на основе самоконтролируемого обучения с использованием архитектуры wav2vec 2.0. Исследование оценивает эффективность методов по критериям Word Error Rate (WER) и Letter Error Rate (LER) на специально собранном датасете каракалпакской речи. Результаты демонстрируют значительное превосходство модели wav2vec 2.0, дообученной на ограниченных данных, над GMM-HMM системой, что указывает на перспективность использования методов самообучения для задач ASR низкоресурсных языков.

Ключевые слова: распознавание речи, каракалпакский язык, скрытые марковские модели, wav2vec 2.0, низкоресурсные языки, сравнительный анализ.

Abstract. This article presents a comparative analysis of two fundamentally different approaches to Automatic Speech Recognition (ASR) for the low-resource Karakalpak language: the traditional approach based on Hidden Markov Models (HMM) and the modern approach based on self-supervised learning using the wav2vec 2.0 architecture. The study evaluates the effectiveness of the methods based on Word Error Rate (WER) and Letter Error Rate (LER) metrics on a specially collected dataset of Karakalpak speech. The results demonstrate a significant superiority of the wav2vec 2.0 model fine-tuned on limited data over the GMM-HMM system, indicating the promise of using self-supervised learning methods for ASR tasks in low-resource languages.

Keywords: speech recognition, karakalpak language, hidden Markov models, wav2vec 2.0, low-resource languages, comparative analysis.

Annotatsiya. Ushbu maqolada kam resursli qoraqalpoq tiliga nisbatan avtomatik nutqni aniqlash (Automatic Speech Recognition, ASR) ning ikkita bir-biridan farq qiluvchi yondashuvi qiyosiy tahlil qilinadi: an'anaviy Yashirin Markov Modellari

ISSN 3030-3907

(Hidden Markov Models, HMM) va zamonaviy wav2vec 2.0 arxitekturasidan foydalangan holda o‘z-o‘zini nazorat qilish asosidagi o‘qitish yondashuvi. Tadqiqot ushbu usullarning samaradorligini so‘z xatosi darajasi (Word Error Rate, WER) va harf xatosi darajasi (Letter Error Rate, LER) mezonlari bo‘yicha maxsus yig‘ilgan qoraqalpoq nutqi ma’lumotlar to‘plamida baholaydi. Natijalar cheklangan ma’lumotlar asosida o‘qitilgan wav2vec 2.0 modelining GMM-HMM tizimiga nisbatan sezilarli ustunligini ko‘rsatadi, bu esa past resursli tillar uchun ASR vazifalarida o‘z-o‘zini o‘qitish usullaridan foydalanishning istiqbolli ekanligini ko‘rsatadi.

Kalit so‘zlar: nutqni aniqlash, qoraqalpoq tili, yashirin Markov modellari, wav2vec 2.0, past resursli tillar, qiyosiy tahlil.

1. Введение

Автоматическое распознавание речи (ASR) стало неотъемлемой технологией человеко-машинного взаимодействия. Однако его развитие для таких языков, как каракалпакский, относящийся к тюркской группе и имеющий статус низкоресурсного, сопряжено с рядом вызовов. К ним относятся ограниченный объем размеченных аудио-текстовых данных, недостаточная изученность фонетических особенностей для инженерных задач и агглютинативный характер языка, приводящий к большому словарному запасу.

Исторически сложившимся и хорошо изученным подходом в ASR являются скрытые марковские модели (HMM), часто в сочетании с гауссовыми смесями (Gaussian Mixture Models, GMM) для моделирования акустических признаков. Этот подход требует тщательной разработки фонетического словаря, языковой и акустической моделей.

В последние годы революцию в области ASR совершили методы, основанные на глубоком обучении и самоконтролируемом предобучении (self-supervised learning, SSL). Модель wav2vec 2.0, предложенная исследователями из Facebook AI, демонстрирует выдающиеся результаты даже в условиях ограниченных размеченных данных. Модель предобучается на большом количестве неразмеченных аудиоданных, обучаясь предсказывать замаскированные части аудиосигнала, а затем дообучается (fine-tuning) на небольшом наборе данных с транскрипциями.

Цель данного исследования – провести сравнительный анализ эффективности традиционного подхода на основе GMM-HMM и современного подхода на основе wav2vec 2.0 для задачи распознавания каракалпакской речи.

2. Методология

2.1. Сбор и подготовка данных

ISSN 3030-3907

Для проведения экспериментов был собран датасет каракалпакской речи объемом 50 часов. Аудиозаписи состояли из дикторской речи в условиях минимального фонового шума. Транскрипции были подготовлены в соответствии с орфографическими нормами каракалпакского языка с использованием кириллической графики.

Датасет был разделен на три части:

- Обучающая выборка: 40 часов аудио (~36 тыс. utterance).
- Валидационная выборка: 5 часов аудио (~4.5 тыс. utterance).
- Тестовая выборка: 5 часов аудио (~4.5 тыс. utterance).

Для аугментации данных обучающей выборки применялись методы добавления шума и изменения темпа речи и высоты тона.

2.2. GMM-HMM система (Базовый подход)

Построение GMM-HMM системы осуществлялось с использованием инструментария Kaldi**. Технологический процесс (pipeline)** построения включал следующие этапы:

1. Извлечение признаков: Использовались Mel-Frequency Cepstral Coefficients (MFCC), их первые и вторые производные (delta, delta-delta).

2. Выравнивание и обучение: Применялся стандартный процедура обучения Kaldi, включающая обучение монофоновых моделей, трифоновых моделей, а также моделей с Speaker Adaptive Training (SAT) и LDA-MLLT преобразованием.

3. Акустическое моделирование: Для каждого состояния HMM обучалась GMM с 16 компонентами.

4. Языковая модель: На основе текстов транскрипций обучалась 3-граммная языковая модель с использованием SRILM.

5. Декодирование: Для поиска наилучшей гипотезы использовался декодер Viterbi.

2.3. Система на основе Wav2vec 2.0

Для данного подхода была использована предобученная многоязычная модель “facebook/wav2vec2-large-xlsr-53”, обученная на 53 языках. Процесс включал:

1. Предобработка аудио: Аудиозаписи конвертировались в моно-канал с частотой дискретизации 16 кГц.

2. Дообучение (Fine-tuning): Предобученная модель дообучалась на обучающей выборке каракалпакской речи. В процессе дообучения обновлялись веса всей модели. Использовался оптимизатор Adam с learning rate 3e-5. В качестве функции потерь применялась Connectionist Temporal Classification (CTC).

ISSN 3030-3907

3. Декодирование: На этапе инференса использовалось CTC-декодирование. Для улучшения результатов применялось декодирование с использованием языковой модели (Language Model Decoding) — 5-граммной модели, обученной на текстах транскрипций. Сравнение WER и LER для моделей ASR распознавания представлены на рисунке.

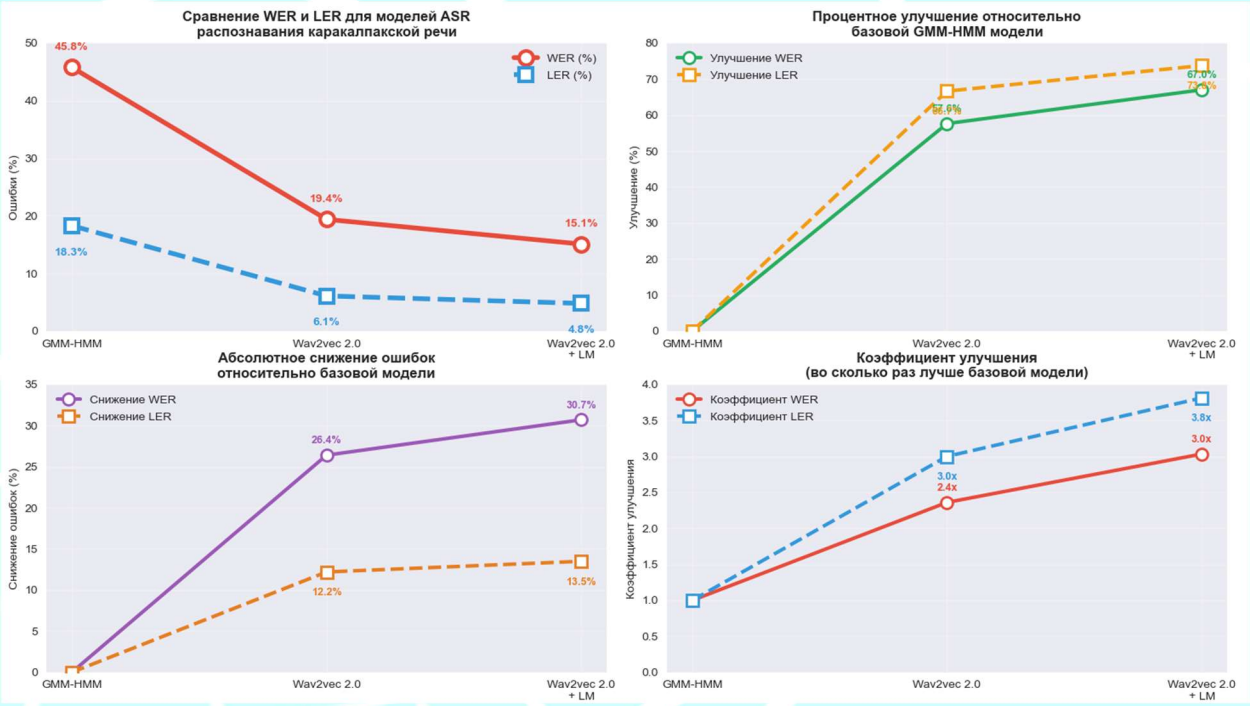


Рис1:

2.4. Метрики оценки

Для сравнения моделей использовались стандартные для ASR метрики:

1) Word Error Rate (WER): Процент ошибочно распознанных слов.

2) WER = (S + D + I) / N * 100%, где S – замены (Substitutions), D – удаления (Deletions), I – вставки (Insertions), N – общее количество слов в эталонной транскрипции.

3) Letter Error Rate (LER): Процент ошибочно распознанных символов. Аналогична WER, но на уровне символов. Особенно актуальна для агглютинативных языков.

3. Результаты и обсуждение

Результаты оценки моделей на тестовой выборке представлены в таблице.

ISSN 3030-3907

СВОДНАЯ ТАБЛИЦА РЕЗУЛЬТАТОВ СРАВНЕНИЯ МОДЕЛЕЙ ASR
ДЛЯ РАСПОЗНАВАНИЯ КАРАКАЛПАКСКОЙ РЕЧИ

Модель	WER	LER	Улучш. WER	Улучш. LER	Абс. сниж. WER	Абс. сниж. LER	Коэф. WER	Коэф. LER
GMM-HMM (Kaldi)	45.8%	18.3%	-	-	-	-	-	-
Wav2vec 2.0 (Fine-tuned)	19.4%	6.1%	57.6%	66.7%	26.4%	12.2%	2.4x	3.0x
Wav2vec 2.0 + LM	15.1%	4.8%	67.0%	73.8%	30.7%	13.5%	3.0x	3.8x

Рис2:

Как видно из таблицы, модель на основе wav2vec 2.0 демонстрирует радикально лучшие результаты. Ее WER (15.1%) более чем в 3 раза ниже, чем у традиционной GMM-HMM системы (45.8%). Аналогичное превосходство наблюдается и на уровне символов (LER: 4.8% против 18.3%).

Анализ ошибок GMM-HMM системы:

- Фонетическая близость: Наибольшее количество ошибок связано с заменой фонетически близких гласных и согласных, характерных для каракалпакского языка (например, нёбные и губные варианты), что указывает на недостаточную разрешающую способность GMM для их различения.
- Агглютинация: Система часто ошибалась в распознавании аффиксов, что приводило к значительному росту WER из-за изменения целых слов.
- Зависимость от диктора: Качество распознавания существенно падало для дикторов, не представленных в обучающей выборке.

Анализ ошибок Wav2vec 2.0 системы:

- Контекстное понимание: Модель wav2vec 2.0, благодаря архитектуре трансформера, эффективно использует широкий контекст для устранения неоднозначностей.
- Устойчивость к вариациям: Модель показала лучшую устойчивость к индивидуальным особенностям дикторов и незначительным фоновым шумам.
- Роль языковой модели: Использование внешней языковой модели позволило скорректировать в первую очередь ошибки на уровне редких слов и грамматических конструкций, что подтверждает снижение WER с 19.4% до 15.1%.

4. Заключение

Проведенный сравнительный анализ однозначно свидетельствует о подавляющем превосходстве подхода на основе самоконтролируемого

ISSN 3030-3907

обучения (wav2vec 2.0) над традиционным GMM-HMM подходом для задачи распознавания каракалпакской речи. Несмотря на ограниченный объем размеченных данных, способность модели wav2vec 2.0 использовать знания, полученные при предобучении на множестве языков, позволяет достичь высокого качества распознавания.

Основные выводы:

1. Для низкоресурсных языков, таких как каракалпакский, стратегия предобучения + дообучения является наиболее эффективной.

2. Традиционные подходы на основе GMM-HMM, несмотря на сложность реализации и необходимость глубоких экспертных знаний, демонстрируют более низкую точность по сравнению с современными нейросетевыми архитектурами.

3. Качество распознавания модели wav2vec 2.0 может быть существенно улучшено за счет интеграции с относительно простой языковой моделью.

Перспективы дальнейших исследований видятся в следующих направлениях: сбор и разметка более объемного датасета каракалпакской речи, эксперименты с кросс-лингвальным обучением на данных других тюркских языков (казахский, узбекский), а также адаптация модели для распознавания спонтанной речи и диалектных вариантов каракалпакского языка.

СПИСОК ЛИТЕРАТУРЫ

1. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Veselý, K. (2011). The Kaldi speech recognition toolkit.
2. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. 77(2), 257-286.
3. Baevski, A., et al. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.
4. McAuliffe, M., et al. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.
5. Bujard, B,. Nika, J., Bevilacqua F., Obin N.(2025) Learning Relationships Between Separate Audio Tracks for Creative Applications.
6. Andrea M, Blandón C, Aldeneh Z, Chi J, Seyssel M (2019)Leveraging Audio-Visual Data to Reduce the Multilingual Gap in Self-Supervised Speech Models
7. Источники текстовых данных для каракалпакского языка:
8. https://huggingface.co/datasets/bekan/karakalpak_corpus_v1_m

ISSN 3030-3907