Bulletin of L.N. Gumilyov ENU. Mathematics. Computer science. Mechanics series, 2023, Vol. 145, №4, P.12-22
http://bulmathmc.enu.kz, E-mail: vest_math@enu.kz

МРНТИ: **28.23.17**

Ж.М. Кожирбаев

National Laboratory Astana, пр. Кабанбай батыр, 53, Астана, Казахстан
(E-mail: zhanibek.kozhirbayev@nu.edu.kz)

Самообучение для улучшения системы распознавания речи на казахском языке [1]

Аннотация: В последнее время достижения в нейронных моделях, обученных с использованием обширных многоязычных текстовых и устных данных, продемонстрировали многообещающий потенциал для улучшения ситуации с языками, которым не хватает ресурсов. Это исследование сосредоточено на проведении экспериментов с использованием передовых моделей распознавания речи, в частности, Wav2Vec2.0 и Wav2Vec2-XLSR, применительно к казахскому языку. Основная цель этого исследования — оценить эффективность этих моделей при расшифровке разговорного казахского содержания. Кроме того, исследование направлено на изучение возможности использования данных из других языков для начального обучения и оценку того, может ли уточнение модели с помощью данных целевого языка повысить ее производительность. Таким образом, это исследование предлагает ценную информацию о жизнеспособности использования предварительно обученных многоязычных моделей в контексте языков с ограниченными ресурсами. Точно настроенная модель wav2vec2.0-XLSR показала исключительные результаты, продемонстрировав коэффициент ошибок символов (CER) 1,9 и коэффициент ошибок слов (WER) 8,9 при сравнении с тестовым набором набора данных kazcorpus. Результаты этого анализа могут способствовать созданию надежных и эффективных систем автоматического распознавания речи (ASR), адаптированных для казахского языка. Эти разработки принесут пользу целому ряду приложений, в том числе преобразованию речи в текст, голосовым помощникам и средствам голосового общения.

Ключевые слова: автоматическое распознавание речи, казахский язык, Wav2Vec 2.0, Wav2Vec2-XLSR, предварительно обученные модели-трансформеры, модели представления речи.

DOI: https://doi.org/10.32523/2616-7182/bulmathenu.2023/4.2

2000 Mathematics Subject Classification: 68T10

1. ВВЕДЕНИЕ

В последнее время модели на основе последовательностей продемонстрировали замечательные достижения в распознавании речи при сопоставлении с традиционными системами автоматического распознавания речи. Эти основанные на последовательности модели используют нейронные сети для преобразования речи в текст, что упрощает процесс моделирования. Среди них архитектура глубоких нейронных сетей Трансформеры [21] выделяется как широко распространенный и продемонстрировает заслуживающие внимания достижения в построении сквозных систем распознавания речи [2–4]. Несмотря на значительный прогресс в развитии моделей ASR, разработка надежных моделей для языков, помимо английского, остается сложной задачей. Эта проблема в основном возникает из-за того, что современные модели, как правило, требуют длительных часов

[1]Работа выполнена при финансовой поддержке Министерства науки и высшего образования Республики Казахстан (проект №AP13068635)

аннотированных речевых данных для обучения, чтобы достичь удовлетворительных результатов. Это особенно верно для казахского, тюркского языка, на котором говорит мировое сообщество, насчитывающее более 13 миллионов человек (согласно статистике веб-сайта Ethnologue[2]).

Недавние успехи в методологиях обучения с самоконтролем продемонстрировали потенциал в решении проблемы ограниченной доступности данных для языков без достаточных ресурсов. Обучение с самостоятельным наблюдением — уникальный подход к совершенствованию систем распознавания речи — использует множество немаркированных речевых данных для получения ценных представлений речевого сигнала. В отличие от традиционного контролируемого обучения, которое требует помеченных данных для обучения модели, алгоритмы самоконтролируемого обучения извлекают знания из необработанных данных без необходимости явных аннотаций. В рамках обучения с самостоятельным наблюдением модели оттачиваются для выполнения задач, тесно связанных с основной задачей распознавания речи, но лишенных необходимости в помеченных входных данных. Метод включает в себя обучение модели прогнозированию последующего кадра речевого сигнала с учетом предыдущих кадров — концепция, называемая контрастным прогностическим кодированием (CPC) [5].

Другая стратегия влечет за собой обучение модели различению двух отдельных речевых сегментов, например, различение между парой близких во времени речевых кадров и теми, которые значительно разнесены друг от друга. Потенциал обучения с самостоятельным наблюдением в сфере распознавания казахской речи заключается в его способности извлекать выгоду из значительных объемов неразмеченных данных, что является значительным преимуществом для языков, где аннотированные данные остаются дефицитными. Используя методы самоконтроля для обучения модели распознавания речи на немаркированных данных, модель может умело расшифровывать ключевые аспекты речевого сигнала, включая фонемы и акустические нюансы, необходимые для точной транскрипции. Помимо своего потенциала для повышения точности систем распознавания речи, обучение с самоконтролем обещает уменьшить количество необходимых помеченных данных для обучения. Это сокращение может существенно сократить затраты и время, необходимые для создания надежной системы распознавания казахской речи, сделав ее более доступной для исследователей и разработчиков. Недавние усовершенствования в кодировщиках звука с самоконтролем, примером которых является Wav2Vec2.0 [6], эффективно ассимилировали высококлассные аудиопредставления. Однако их неконтролируемый подход к предварительному обучению создает проблему эффективного преобразования этих представлений в практические результаты. Следовательно, фаза тонкой настройки становится обязательной для профессионального развертывания этих моделей для таких задач, как автоматическое распознавание речи. Существуют две опции предварительно обученных моделей Wav2Vec2.0: первая предварительно обучена исключительно для одного языка, а вторая — многоязычная предварительно обученная модель (XLSR-53). Это исследование было предпринято с целью противопоставления этих двух подходов, чтобы установить их эффективность в обеспечении надежного ASR для казахского языка. Основные вклады этого исследования включают:

1. В дополнение к имеющимся речевым корпусам на казахском языке мы собрали аудиозаписи вместе с соответствующими им текстами из общедоступных источников. Совокупные данные, собранные примерно за 1000 часов. Каждому аудиофайлу соответствовал отдельный текстовый файл, содержащий содержание аудиокниги. Примечательно, что необходимо признать, что между звуком и текстом отсутствовала синхронизация, что подразумевает отсутствие выравнивания на уровне предложения или слова. Чтобы решить эту проблему рассогласования, мы использовали технику сегментации, основанную на алгоритме коннекционистской временной классификации (CTC) [7]. Этот подход облегчил точное извлечение выравнивания аудио-текста.

[2]https://www.ethnologue.com/language/kaz

2. Была проведена серия экспериментов с использованием базовой архитектуры Wav2Vec2.0 и XLSR-53, охватывающих различные сценарии предварительной подготовки и тонкой настройки.
3. Был проведен обширный сравнительный анализ двух методологий: Wav2Vec2.0 и Wav2Vec2.0-XLSR.

2. ОСНОВНАЯ ЧАСТЬ

В этом разделе представлен краткий обзор соответствующей литературы, связанной с этой статьей, разделенный на два подраздела: Wav2Vec2.0 и XLSR-53, а также Казахский ASR.

Wav2Vec 2.0. тщательно разработан для расшифровки речи, встроенной в аудиосигналы, с использованием самоконтролируемой методологии предварительного обучения, которая впитывает идеи из огромных объемов немаркированных аудиоданных. Он объединяет принципы нескольких предшествующих моделей, а именно, Contrastive Predictive Coding (CPC) [5], Model Predictive Control (MPC) [8], wav2vec [9] и vqwav2vec [10]. Архитектура Wav2Vec2 гармонизирует сверточные нейронные сети (CNN) и преобразователи, что позволяет ему воспринимать как локальные нюансы, так и всеобъемлющие шаблоны в аудиоданных. В модели используется многоуровневый сверточный кодировщик признаков, обозначенный как $f : X \to Z$, для кодирования необработанных аудиосигналов X в представления скрытой речи $z_1, \ldots, z_T$, которые затем подаются в сеть с масками преобразователя, обозначаемые как $g : Z \to C$, который отображает представления из скрытого пространства в дискретный набор выходных данных, $q_1, \ldots, q_T$, которые представляют цели в самоконтролируемой цели обучения [6, 12]. Модуль преобразования контекстуализирует квантованные представления с помощью блоков внимания, в результате чего получается набор дискретных контекстуальных представлений $c_1, \ldots, c_T$. Кодировщик функций состоит из семи сверточных блоков, каждый из которых имеет 512 каналов, ширину ядра {10, 3, 3, 3, 3, 2, 2} и шаги {5, 2, 2, 2, 2, 2, 2}. С другой стороны, трансформаторная сеть состоит из 24 блоков с 1024 размерами и 4096 внутренними размерами. Всего в ней также имеется 16 головок внимания.

XLSR-53 [11] представляет собой многоязычную модель, основанную на межъязыковой модели XLM-R, предназначенную для решения задач многоязычной и межъязыковой обработки естественного языка (NLP). Основываясь на модели Wav2Vec 2.0, XLSR-53 обладает способностью получать скрытое квантование, охватывающее различные языки. Это достигается за счет использования квантования произведения для отбора квантованных представлений из кодовых книг. В процессе отбора используется метод Gumbel-Softmax, обеспечивающий полное различие. Архитектура XLSR-53 похожа на архитектуру двунаправленный кодировщик представлений трансформера (BERT) [12] с заметным отличием: она включает в себя 53 языковых вложения для каждого из поддерживаемых языков. Этот сложный дизайн позволяет модели обрабатывать различные языки, улавливая их тонкости даже в случаях схожего написания или произношения. Кроме того, XLSR-53 может похвастаться огромным количеством параметров, составляющим 500 миллионов, что делает его одной из крупнейших доступных многоязычных моделей. Эта модель обучается на обширном и разнообразном корпусе, включающем текстовые данные речи, извлеченные из более чем 53 языков. Присущая XLSR-53 способность понимать несколько языков делает его исключительно выгодным для межъязыкового трансферного обучения. Это влечет за собой адаптацию модели, обученной на одном языке, для хорошей работы на другом языке, требуя лишь минимального дополнительного обучения.

Недавние достижения в области ASR открыли новые сквозные архитектуры, демонстрирующие впечатляющую точность при наличии достаточного количества наборов данных. Основной принцип, лежащий в основе этих сквозных моделей, вращается вокруг прямого преобразования входных речевых сигналов в последовательности символов. Такой оптимизированный подход оптимизирует процедуры обучения,

тонкой настройки и логического вывода. В области исследований ASR эксперты преимущественно тяготеют к двум различным методологиям обучения систем ASR: полностью контролируемым и самоконтролируемым моделям. В контексте первой категории Есенбаев [13] провели обширное исследование, направленное на преодоление проблемы автоматического, независимого от говорящего распознавания слитной казахской речи, сосредоточив внимание на определенной лексической основе в условиях шумной среды. Предложенная авторами система продемонстрировала похвальные результаты в различных задачах, включая фонетическое распознавание английской речи, а также распознавание непрерывной казахской речи. Примечательно, что система продемонстрировала относительное улучшение качества распознавания до 20%. Обращает на себя внимание достижение качества распознавания 94,5% для казахской речи. По сути, это исследование выступает в качестве основополагающего шага, закладывающего основу для последующей разработки более продвинутых платформ, предназначенных для непрерывного распознавания казахской речи. В своем исследовании авторы [14] углубляются в распознавание потоковой речи посредством реализации модели RNN-T. Эта архитектура построена с использованием нейронных сетей, таких как LSTM и BLSTM, с использованием обучающего набора данных продолжительностью более 300 часов, включающего как подготовленные (чтение), так и записи спонтанной речи. Результаты исследования подчеркивают способность модели RNN-T достигать CER 10,6. В отдельном исследовании тех же авторов Мамырбаев [15] представляет гибридную модель Transformer + CTC (Connectionist Temporal Classification). Он был отточен с использованием набора речевых данных, охватывающего 400 часов. Примечательно, что результаты исследования подчеркивают эффективность модели, регистрируя CER 3,7 и WER 8,3. Безусловно, следует отметить совместную работу исследователей Центра речевых технологий Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики совместно с Костанайским государственным университетом им. А. Байтурсынова [16]. Эти исследователи приступили к проекту, сосредоточенному на признании и синтезе двуязычного (казахско-русского) языка. Их усилия были сосредоточены на развитии области обработки двуязычной речи, тем самым способствуя более широкому спектру исследований в области языковых технологий. Авторы [17] внесли значительный вклад в исследование распознавания речи на казахском языке в виде базы данных KSC 1. Эта обширная база данных служит открытым эталоном, содержащим около 332 часов расшифрованного аудио. База данных включает более 153 000 высказываний, произнесенных людьми из разных возрастных групп, регионов и полов. Авторы использовали сквозную модель (E2E) на основе Transformer, достигнув коэффициента ошибок символов (CER) 2,8 и коэффициента ошибок слов (WER) 8,7 в этом наборе данных. В родственной разработке Муссаходжаева [18] расширили базу данных KSC до ошеломляющих 1128 часов. Это расширение включало в себя включение дополнительных данных из различных источников, включая телевизионные новости, теле- и радиопрограммы, парламентские выступления и подкасты. Авторы тщательно определили спецификации корпуса и обосновали его полезность, используя модель ASR на основе Transformer. Эта модель дала многообещающие результаты: общая частота ошибок в словах (WER) составила 15,1 в проверочном наборе и 15,6 в тестовом наборе. Эти усилия значительно расширяют ресурсы, доступные для развития технологии распознавания речи на казахском языке. Хотя продолжающаяся разработка моделей продемонстрировала выдающееся мастерство, значительная часть из них в значительной степени зависит от методологий обучения с учителем, что требует значительных объемов аннотированных данных. К сожалению, процесс маркировки и аннотирования данных является ресурсоемким, дорогостоящим и трудоемким, часто влекущим за собой ручное вмешательство. Более того, могут возникнуть обстоятельства, когда получение таких данных становится нецелесообразным из-за ограничений или отсутствия. В отличие от области полностью контролируемых моделей, недавние исследования были направлены на использование существенных акустических моделей, обученных с помощью методов самоконтролируемого обучения и обширных резервуаров немаркированных

данных. В качестве примера можно привести работу [19], в которой представили неконтролируемое предварительное обучение с использованием Wav2Vec2.0. Авторы интегрировали факторизованный уровень TDNN, чтобы поддерживать связь между голосом и временными шагами, тем самым повышая эффективность распознавания речи для казахского языка. Кроме того, они использовали многоязычную предварительную подготовку и методы синтеза речи для дальнейшего повышения производительности. Результаты их экспериментов подчеркнули преимущества ассимиляции немаркированных данных из языков за пределами целевого и использования улучшения данных посредством синтеза речи. Эти подходы, в частности, привели к существенному снижению частоты ошибок в словах в наборах тестов. Это исследование знаменует собой заметный шаг в направлении оптимизации систем распознавания речи при одновременном снижении зависимости от обширных помеченных наборов данных.

3. МЕТОДОЛОГИЯ

Этот раздел сосредоточен на наборах данных, специально предназначенных для распознавания речи на казахском языке. Он также углубляется в методологии, используемые для разработки модулей распознавания речи, предназначенных для этого языкового контекста.

ISSAI KSC. Набор данных ISSAI KSC является наиболее обширным открытым ресурсом, созданным для поддержки казахстанских приложений обработки речи и языка [17]. Этот существенный набор данных включает более 332 часов контента, собранного через веб-платформу, предназначенную для записи речи. Эта платформа приглашала добровольцев формулировать предложения, взятые из целого ряда источников, включая книги, законы, Википедию, новостные порталы и блоги. Набор данных KSC может похвастаться разнообразием, включая динамики и аудиозаписи из разных регионов Казахстана с использованием различных устройств, таких как смартфоны, планшеты и ноутбуки. Пул спикеров родом из пяти разных регионов, при этом наборы для проверки и тестирования включают 51,7% спикеров-женщин и 48,3% спикеров-мужчин.

Kazcorpus. Акустический корпус kazcorpus [20] включает в себя два отдельных подкорпуса: kazspeechdb и kazmedia. На основе корпуса kazspeechdb была заложена основа для построения корпуса новостей вещания. Этот подкорпус состоит из речевых фрагментов, а именно 12 675 предложений на казахском языке, записанных в контролируемых студийных условиях. Спикеры принадлежат к разным полам, возрастам и регионам Казахстана. Подкорпус охватывает в общей сложности 22 часа речи с участием 169 говорящих, включая 73 мужских и 96 женских голосов. Каждый спикер произнес по 75 предложений.

С другой стороны, подкорпус KazMedia включает в себя аудио- и текстовые данные, взятые с официальных сайтов телеинформационных агентств, в частности «Хабар», «Астана ТВ» и «31 канал». Текстовые данные представляют собой совокупность новостных статей на казахском языке, публикуемых на официальных сайтах этих каналов. Аудиоданные состоят из файлов WAV, которые представляют собой звуковые дорожки, извлеченные из различных сегментов видеоновостей, транслируемых на этих каналах на казахском языке. В совокупности этот подкорпус включает 21 час речи.

KazLibriSpeech. Мы собрали аудиозаписи в сочетании с соответствующими текстами из открытых источников, в общей сложности около 1000 часов данных. Каждый аудиофайл выравнивается с соответствующим текстом в аудиокниге, хотя выравнивания на уровне предложения или слова нет. Следовательно, последующая попытка состоит в том, чтобы разделить эти аудиофайлы на более мелкие интервалы, будь то слова, фразы или предложения. Затем каждый такой сегмент необходимо сравнить с соответствующим ему озвученным текстом, воспроизведенным в том же интервале. Хотя процесс выравнивания и сегментации может быть сложным, этот метод позволяет создавать существенные наборы данных, охватывающие различные источники и домены, с минимальными затратами.

Учитывая, что качество собранных аудиокниг варьируется, был начат процесс очистки и нормализации. Это включало удаление шума, управление омоглифами, транслитерацию, извлечение невоспроизводимых фрагментов, замену акронимов и аббревиатур их полными формами, числовую нормализацию, замену символов их фонетическими аналогами, первоначальную сегментацию на уровне глав и разделение текста. Преобразование исходного текста в краткие предложения с использованием знаков препинания. Кроме того, любые сопутствующие музыкальные элементы в начале и конце аудиокниг были удалены.

Наш подход к сегментации заключался в использовании алгоритма CTC, который обеспечивает точное выравнивание аудиотекста, даже если аудиозапись включает неразборчивые речевые фрагменты в начале или в конце. Наш метод включает в себя обучение сквозной сети на предварительно выровненных данных с использованием системы ASR на основе CTC-внимания. CTC, как механизм вывода и оценки нейронной сети, играет важную роль в обучении рекуррентных нейронных сетей для решения задач, основанных на последовательности, с учетом переменного времени. Этот механизм не зависит от базовой структуры нейронной сети. В нашем контексте эта модель разграничивает речевые сегменты в аудиофайлах на уровне предложений. Модель ASR, необходимая для сегментации, была отточена с использованием набора данных ISSAI KSC в инструменте Espnet [21].

Более подробное описание имеющихся корпусов, предназначенных для распознавания речи на казахском языке, приведено в Таблице 1.

Таблица 1 – Структура корпусов казахского языка

Структура	**Наименование корпуса/наборов**	**Тип данных**	**Количество wav-файлов**	**Общая продолжительность wav-файлов (час)**
1	ISSAI KSC	Краудсорсинговые записи	153853	332.6
1.1	Train		147236	318.4
1.2	Dev		3283	7.1
1.3	Test		3334	7.1
2	Kazcorpus	Смешанный тип: студийные записи, подготовленная речь + спонтанная речь в разных акустических условиях	13425	44.16
2.1	kazspeechdb		12675	22.61
2.1.1	Train		11175	19.92
2.1.2	Dev		750	1.36
2.1.3	Test		750	1.34
2.2	KazMedia		740	21.55
2.2.1	Train		561	18.04
2.2.2	Dev		49	1.00
2.2.3	Test		130	2.51
3	KazLibriSpeech	Аудиокниги	575243	992

В контексте данного исследования была проведена оценка двух вариантов wav2vec: (1) Wav2Vec2.0, предварительно обученный и настроенный исключительно для казахского языка, и (2) XLSR-53, изначально предварительно обученный для 53 языка с последующим непрерывным предварительным обучением и доводкой на казахский язык.

Wav2Vec 2.0. Эксперимент проводился с использованием платформы Fairseq. Базовая модель Wav2Vec 2.0 прошла предварительное обучение с использованием различных конфигураций, включающих в себя уровень отбрасывания кодировщика, установленный на 0,05, dropout_input, dropout_features и feature_grad_mult, установленный на 0,1, и encoder_embed_dim, установленный на 768. Гиперпараметры обучения включали скорость обучения $5 * 10^{-4}$, с фазой разминки в первые 10% продолжительности тренировки. Количество обновлений было указано как 800 000, а максимальное количество токенов — 1 200 000. Последовательно применялся оптимизатор Адама в соответствии с исходным подходом.

Для тонкой настройки использовались стандартные процедуры с параметрами, заданными следующим образом: количество обновлений достигло 160 000, а максимальное

количество токенов составило 2 800 000. Как и в случае с предварительным обучением, был задействован оптимизатор Adam, использующий скорость обучения $3 * 10^{-5}$ и накопление градиента из 12 шагов. Размер пакета обучения динамически определялся платформой с учетом заданного максимума маркеров. В ходе обучения оптимальная модель выбиралась на основе наименьшего значения WER, достигнутого в проверочном наборе.

XLSR-53. Модель XLSR прошла предварительную подготовку с идентичными конфигурациями, используемыми для большой модели Wav2Vec. Блок кодера состоял из 24 слоев, каждый размером 1024, и использовался набор из 16 блоков внимания без учета исключения. Параметры тонкой настройки были определены в соответствии с конфигурациями, примененными в исходном эксперименте XLSR с Wav2Vec 2.0.

Языковая модель. После процесса тонкой настройки модель подвергается декодированию с помощью 3-граммовой языковой модели. Языковая модель была обучена с использованием транскрипций из всех доступных наборов данных, указанных в таблице 1, с использованием набора инструментов Kenlm. Для целей декодирования используется декодер поиска луча с размером луча, настроенным на 1500.

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

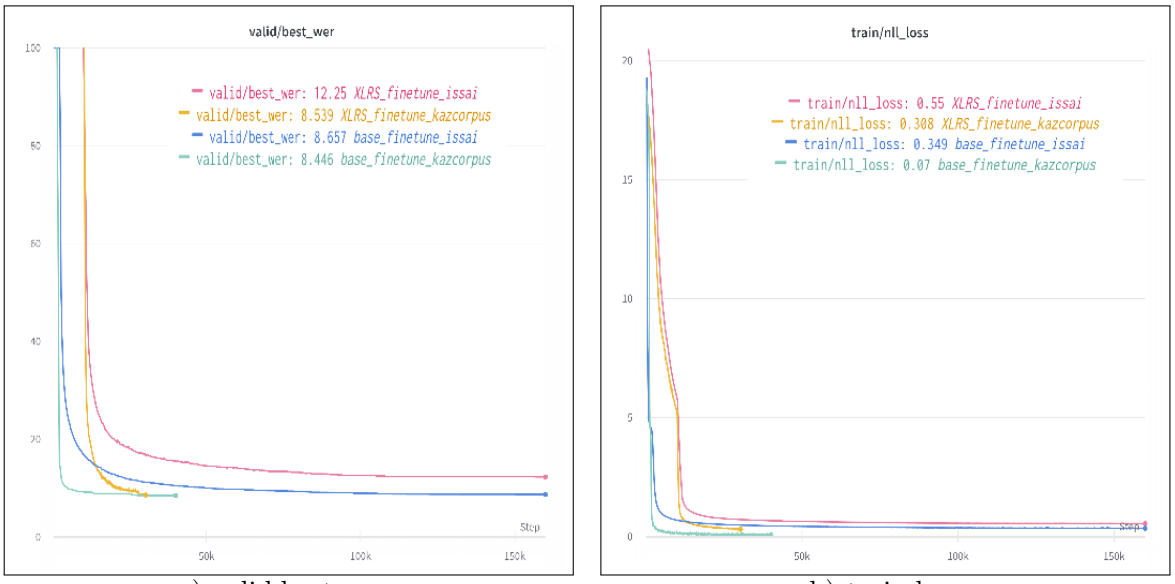
При оценке моделей Wav2Vec2.0 использовались наборы данных, описанные в разделе 3. Производительность систем ASR, указанная в баллах WER и CER, представлена в таблице 2. Для каждой архитектуры использовались различные сценарии обучения с разными параметрами. Эксперименты проводились на сервере NVIDIA DGX-1, оснащенном 8 графическими процессорами V100.

Таблица 2 – Производительность моделей Wav2Vec 2.0

ID	**Исходная модель**	**Набор данных предварительной подготовки**	**Набор данных тонкой настройки**	**Оценочный набор**	**Набор ЯМ**	**Тест WER**	**Тест CER**
1	Wav2Vec 2.0 Base	KazLibriSpeech	Kazcorpus (train+dev)	Kazcorpus (test)	3-gram KenLM	12.4	2.6
2	Wav2Vec 2.0 Base	KazLibriSpeech	ISSAI KSC1 (train+dev)	ISSAI KSC1 (test)	3-gram KenLM	10.1	2.8
3	XLSR-53	KazLibriSpeech	Kazcorpus (train+dev)	Kazcorpus (test)	3-gram KenLM	8.9	1.9
4	XLSR-53	KazLibriSpeech	ISSAI KSC1 (train+dev)	ISSAI KSC1 (test)	3-gram KenLM	15.1	4.8

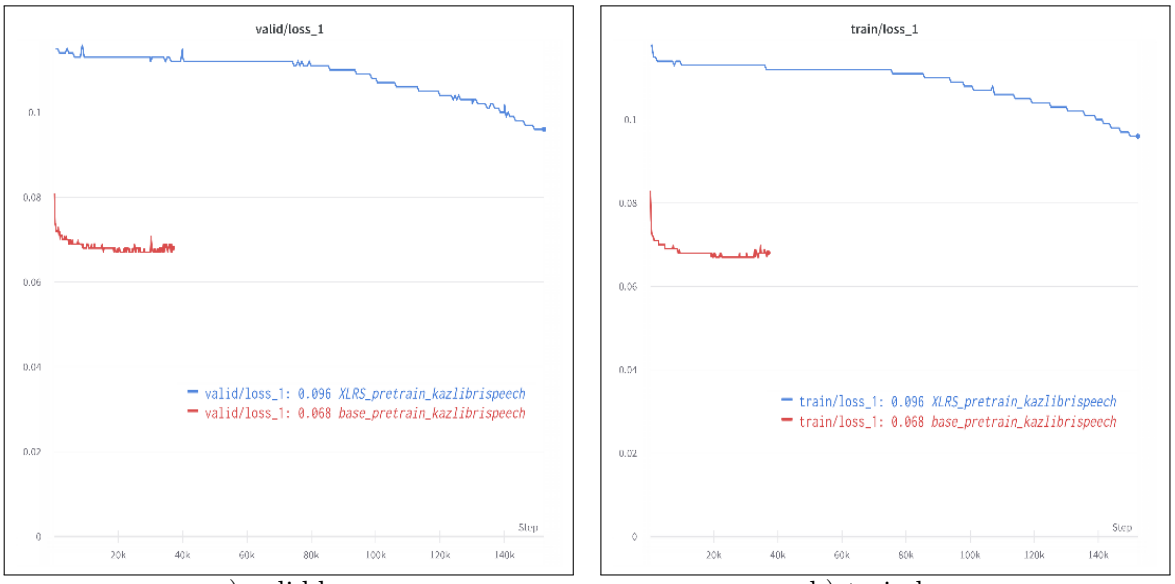
В Таблице 2 показаны баллы частоты ошибок символов и ошибок слов для точно настроенных моделей Wav2Vec 2.0-base и XLSR-53. На этапе предварительного обучения использовался исключительно корпус KazLibriSpeech, а для тонкой настройки использовались ISSAI KSC1 и Kazcorpus. Результаты подчеркивают исключительную производительность предварительно обученной модели XLSR-53. После предварительной подготовки с использованием корпуса KazLi-briSpeech и тонкой настройки с использованием данных Kazcorpus (train+dev) он достигает CER 1,9 и WER 8,9 на тестовом наборе. Эти цифры заметно выше на 28,2% и 26,9% по сравнению с базовой моделью Wav2Vec 2.0 с точки зрения WER и CER соответственно. Эти результаты подчеркивают значительное улучшение производительности модели благодаря предварительному обучению, где размер набора данных, используемого для предварительного обучения, играет ключевую роль. И наоборот, базовая модель Wav2Vec 2.0, прошедшая предварительное обучение с помощью корпуса KazLibriSpeech и тонкую

настройку с использованием данных ISSAI KSC1 (train+dev), демонстрирует превосходные результаты по сравнению с моделью XLSR-53. Это событие могло быть вызвано тем, что исходная модель прошла предварительное обучение с использованием набора данных, содержащего аудиокниги. Ход обучения на этапах предварительного обучения и тонкой настройки показан на Рисунках 1 и 2. Эти визуальные эффекты отображают изменения значений потерь в ходе обучения, предлагая представление о процессе оптимизации моделей.



a) valid:best_wer b) train:loss

Рисунок 1 – Точная настройка Wav2Vec 2.0



a) valid:loss b) train:loss

Рисунок 2 – Предварительное обучение Wav2Vec 2.0

5. ВЫВОДЫ

Основная цель этого исследования — оценить эффективность передовых моделей распознавания речи при расшифровке казахского языка, который относится к категории

малоресурсных языков. Наш анализ включает в себя всестороннее сравнение этих моделей с учетом характера и объема данных, используемых на этапах предварительной подготовки и тонкой настройки. Кроме того, исследование направлено на раскрытие потенциальных преимуществ первоначального предварительного обучения с использованием данных из других языков с последующей тонкой настройкой данных из целевого языка. Исследование включает в себя серию экспериментов с использованием архитектур Wav2Vec2.0 и Wav2Vec2-XLS-R, в которых исследуются различные сценарии предварительной подготовки и тонкой настройки. С помощью этих экспериментов мы не только получаем представление о производительности этих моделей специально для казахского языка, но и выясняем последствия, применимые к другим языкам и условиям.

По сути, это исследование освещает перспективы использования усовершенствованных многоязычных моделей и сопоставления самоконтролируемых и полностью контролируемых методов для надежного автоматического распознавания речи в языковых контекстах с ограниченными ресурсами. Выводы и методологии, представленные в этом исследовании, имеют более широкое значение, распространяющееся на смягчение ограничений языковых ресурсов и продвижение разработки системы ASR для широкого круга языков.

Список литературы

1 Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I. Attention is all you need // Advances in neural information processing systems. - 2017. - V. 30. - P. 6000–6010.

2 Karita S., Chen N., Hayashi T., Hori T., Inaguma H., Jiang Z., Someki M., Soplin N., Yamamoto R., Wang X., Watanabe S. A comparative study on transformer vs rnn in speech applications // Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). - Singapore, 2019. - P. 449-456.

3 Nakatani T. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration // Proceedings of Interspeech. - Graz, Austria, 2019. - P. 1408-1412.

4 Dong L., Xu S., Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition // Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP). - Calgary, Canada, 2018. - P. 5884-5888.

5 Song J., Ermon S. Multi-label contrastive predictive coding // Advances in Neural Information Processing Systems. - 2020. - V. 33. - P. 8161-8173.

6 Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in neural information processing systems. - 2020. - V. 33. - P. 12449-12460.

7 K?rzinger L., Winkelbauer D., Li L., Watzel T., Rigoll G. CTC-segmentation of large corpora for german end-to-end speech recognition // Proceedings of Speech and Computer: 22nd International Conference (SPECOM). - St. Petersburg, Russia, 2020. - P. 267-278.

8 Li S., Li L., Hong Q., Liu L. Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning // Proceedings of Interspeech. - Shanghai, China, 2020. - P. 5006-5010.

9 Schneider S., Baevski A., Collobert R., Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition // Proceedings of Interspeech. - Graz, Austria, 2019. - P. 3465-3469.

10 Baevski A., Schneider S., Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations // Proceedings of 8th International Conference on Learning Representations (ICLR). - Addis Ababa, Ethiopia, 2020. - P. 1-12.

11 Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. Unsupervised cross-lingual representation learning for speech recognition // Proceedings of Interspeech. - Brno, Czechia, 2021. - P. 2426-2430.

12 Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. - Minneapolis, Minnesota, USA, 2019. - P. 4171-4186.

13 Yessenbayev Z., Karabalayeva M., Shamayeva F. Large Vocabulary Continuous Speech Recognition for Kazakh // Proceedings of the I International Conference on Computer processing of Turkic Languages. - Astana, Kazakhstan, 2013. - P. 217-221.

14 Mamyrbayev O., Oralbekova D., Kydyrbekova A., Turdalykyzy T., Bekarystankyzy A. End-to-end model based on RNN-T for Kazakh speech recognition // Proceedings of the 3rd International Conference on Computer Communication and the Internet (ICCCI). - Nagoya, Japan, 2021. - P. 163-167.

15 Mamyrbayev O., Oralbekova D., Alimhan K., Nuranbayeva B.Hybrid end-to-end model for Kazakh speech recognition // International Journal of Speech Technology. - 2022. - P. 1-10.

16 Khomitsevich O., Mendelev V., Tomashenko N., Rybin S., Medennikov I., Kudubayeva S. A blingual Kazakh-Russian system for automatic speech recognition and synthesis // Proceedings of Speech and Computer: 17th International Conference (SPECOM). - Athens, Greece, 2015. - P. 25-33.
17 K7assanov Y., Mussakhojayeva S., Mirzakhmetov A., Adiyev A., Nurpeiissov M., Varol H. A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. - Virtual, 2021. - P. 697-706.
18 Mussakhojayeva S., Khassanov Y., Varol H. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus // Proceedings of Interspeech. - Incheon, Republic of Korea, 2015. - P. 18-22.
19 Meng W., Yolwas N. A Study of Speech Recognition for Kazakh Based on Unsupervised Pre-Training // Sensors. - 2023. - Vol. 23, №2. - P. 870-883.
20 Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. Assembling the kazakh language corpus // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. - Seattle, Washington, USA, 2013. - P. 1022-103.
21 Watanabe S., Hori T., Karita S., Hayashi T., Nishitoba J., Unno Y., Soplin N., Heymann J., Wiesner M., Chen N., Renduchintala A. Espnet: End-to-end speech processing toolkit // Proceedings of Interspeech. - Hyderabad, India, 2018. - P. 2207-2211.

Ж.М. Кожирбаев

National Laboratory Astana, Қабанбай батыр даңғылы, 53, Астана, Қазақстан

Қазақ тілін тану жүйесін өз бетінше оқыту

Аннотация: Жақында ауқымды көптілді мәтіндік және ауызша деректерге үйретілген нейрондық модельдердегі жетістіктер ресурсы аз тілдердің жағдайын жақсартудың перспективалы әлеуетін көрсетті. Бұл зерттеу сөзді танудың жетілдірілген үлгілерін, атап айтқанда Wav2Vec2.0 және Wav2Vec2-XLSR, қазақ тіліне қолданылатын эксперименттерді жүргізуге бағытталған. Бұл зерттеудің негізгі мақсаты – осы үлгілердің ауызша қазақша мазмұнды транскрипциялаудағы тиімділігін бағалау. Сонымен қатар, зерттеу бастапқы оқыту үшін басқа тілдердегі деректерді пайдалану мүмкіндігін зерттеуге және үлгіні мақсатты тілдегі деректермен нақтылау оның өнімділігін жақсартуға болатынын бағалауға бағытталған. Осылайша, бұл зерттеу ресурс шектеулі тілдер контекстінде алдын ала дайындалған көптілді модельдерді пайдаланудың өміршеңдігі туралы құнды ақпаратты ұсынады. Жақсы бапталған wav2vec2.0-XLSR моделі kazcorpus деректер жинағының сынақ жинағымен салыстырғанда 1,9 таңба қатесінің көрсеткісін (CER) және 8,9 сөз қатесінің көрсеткісін (WER) бере отырып, өте жақсы жұмыс жасады. Бұл талдаудың нәтижелері қазақ тіліне бейімделген сенімді және тиімді сөзді автоматты түрде тану (ASR) жүйелерін құруға ықпал ете алады. Бұл әзірлемелер сөйлеуден мәтінге, дауыстық көмекшілер мен дауыстық байланыстарды қоса алғанда, бірқатар қолданбаларға пайдалы болады.

Түйін сөздер: автоматты түрде сөйлеуді тану, қазақ тілі, Wav2Vec 2.0, Wav2Vec2-XLSR, алдын ала дайындалған трансформер үлгілері, сөйлеуді бейнелеу үлгілері.

Zh.M. Kozhirbayev

National Laboratory Astana, Kabanbay batyr ave. 53, Astana, Kazakhstan

Self-Supervised Training for the Kazakh Speech Recognition System

Abstract: In recent times, advancements in neural models trained using extensive multilingual textual and spoken data have displayed promising potential for enhancing the situation of languages that lack resources. This study is centered on conducting experiments utilizing cutting-edge speech recognition models, specifically Wav2Vec2.0 and Wav2Vec2-XLSR, applied to the Kazakh language. The primary aim of this research is to assess the efficacy of these models in transcribing spoken Kazakh content. Additionally, the investigation seeks to explore the feasibility of leveraging data from other languages for initial training, and to assess whether refining the model with target language data can enhance its performance. As such, this study offers valuable insights into the viability of employing pre-trained multilingual models in the context of underresourced languages. The fine-tuned wav2vec2.0-XLSR model achieved exceptional results, boasting a character error rate (CER) of 1.9 and a word error rate (WER) of 8.9 when evaluated against the test set of the kazcorpus dataset. The outcomes of this analysis hold potential to advance the creation of robust and efficient Automatic Speech Recognition (ASR) systems tailored for the Kazakh language. These developments stand to benefit a range of applications, including speech-to-text translation, voice-activated assistants, and speech-driven communication tools.

Keywords: automatic speech recognition, Kazakh language, Wav2Vec 2.0, Wav2Vec2-XLSR, pre-trained transformer models, speech representation models.

References

1 Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I. Attention is all you need [Advances in neural information processing systems]. 2017. Vol. 30. P. 6000–6010.
2 Karita S., Chen N., Hayashi T., Hori T., Inaguma H., Jiang Z., Someki M., Soplin N., Yamamoto R., Wang X., Watanabe S. A comparative study on transformer vs rnn in speech applications [Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)], Singapore, 2019. P. 449-456.

3 Nakatani T. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration [Proceedings of Interspeech], Graz, Austria, 2019. P. 1408-1412.
4 Dong L., Xu S., Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition [Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP)], Calgary, Canada, 2018. P. 5884-5888.
5 Song J., Ermon S. Multi-label contrastive predictive coding [Advances in Neural Information Processing Systems]. 2020. Vol. 33. P. 8161-8173.
6 Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations [Advances in neural information processing systems]. 2020. Vol. 33. P. 12449-12460.
7 K?rzinger L., Winkelbauer D., Li L., Watzel T., Rigoll G. CTC-segmentation of large corpora for german end-to-end speech recognition [Proceedings of Speech and Computer: 22nd International Conference (SPECOM)], St. Petersburg, Russia, 2020. P. 267-278.
8 Li S., Li L., Hong Q., Liu L. Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning [Proceedings of Interspeech], Shanghai, China, 2020. P. 5006-5010.
9 Schneider S., Baevski A., Collobert R., Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition [Proceedings of Interspeech], Graz, Austria, 2019. P. 3465-3469.
10 Baevski A., Schneider S., Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations [Proceedings of 8th International Conference on Learning Representations (ICLR)], Addis Ababa, Ethiopia, 2020. P. 1-12.
11 Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. Unsupervised cross-lingual representation learning for speech recognition [Proceedings of Interspeech], Brno, Czechia, 2021. P. 2426-2430.
12 Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies], Minneapolis, Minnesota, USA, 2019. P. 4171-4186.
13 Yessenbayev Z., Karabalayeva M., Shamayeva F. Large Vocabulary Continuous Speech Recognition for Kazakh [Proceedings of the I International Conference on Computer processing of Turkic Languages], Astana, Kazakhstan, 2013. P. 217-221.
14 Mamyrbayev O., Oralbekova D., Kydyrbekova A., Turdalykyzy T., Bekarystankyzy A. End-to-end model based on RNN-T for Kazakh speech recognition [Proceedings of the 3rd International Conference on Computer Communication and the Internet (ICCCI)], Nagoya, Japan, 2021. P. 163-167.
15 Mamyrbayev O., Oralbekova D., Alimhan K., Nuranbayeva B.Hybrid end-to-end model for Kazakh speech recognition [International Journal of Speech Technology]. 2022. P. 1-10.
16 Khomitsevich O., Mendelev V., Tomashenko N., Rybin S., Medennikov I., Kudubayeva S. A bilingual Kazakh-Russian system for automatic speech recognition and synthesis [Proceedings of Speech and Computer: 17th International Conference (SPECOM)], Athens, Greece, 2015. P. 25-33.
17 Khassanov Y., Mussakhojayeva S., Mirzakhmetov A., Adiyev A., Nurpeiissov M., Varol H. A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline [Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume], Virtual, 2021. P. 697-706.
18 Mussakhojayeva S., Khassanov Y., Varol H. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus [Proceedings of Interspeech], Incheon, Republic of Korea, 2015. P. 18-22.
19 Meng W., Yolwas N. A Study of Speech Recognition for Kazakh Based on Unsupervised Pre-Training [Sensors]. 2023. Vol. 23. №2, P. 870-883.
20 Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. Assembling the kazakh language corpus [Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing], Seattle, Washington, USA, 2013. P. 1022-103.
21 Watanabe S., Hori T., Karita S., Hayashi T., Nishitoba J., Unno Y., Soplin N., Heymann J., Wiesner M., Chen N., Renduchintala A. Espnet: End-to-end speech processing toolkit [Proceedings of Interspeech], Hyderabad, India, 2018. P. 2207-2211.

Сведения об авторах:

Кожирбаев Ж.М. – PhD, Старший научный сотрудник, National Laboratory Astana, пр. Кабанбай батыр, 53, Астана, Казахстан.

Kozhirbayev Zh. M. – PhD, Senior Researcher, National Laboratory Astana, Kabanbay batyr ave. 53, Astana, Kazakhstan.

Поступила в редакцию 13.09.2023