Чимитов Бэлиг Баирович✉, **Завьялова Виктория Львовна**
Дальневосточный федеральный университет
г. Владивосток, Российская Федерация
chimitov.bb@dvmfu.ru

Дикторозависимое автоматическое распознавание речи на бурятском языке

Аннотация

На фоне доминирования русского языка в сферах образования, делового общения, средствах массовой информации потребность в родном бурятском языке постепенно уменьшается, поэтому актуальный вопрос сохранения языка имеет особое значение для этнического населения. В современных условиях для решения данной проблемы особое внимание уделяется речевым технологиям, основанным на естественной обработке речи, искусственном интеллекте и других методах в целях цифровизации бурятского языка и обеспечения доступа к его ресурсам в электронных средах. Анализ современных ресурсов позволяет сделать вывод об отсутствии технологий, ориентированных на бурятский язык и одновременно позволяющих работать со звуковыми данными. Цель статьи – представить разработку прототипа модели распознавания речи на бурятском языке для автоматической обработки речевых данных. В модели используется архитектура DeepSpeech2, которая позволяет сочетать функции свёрточных нейронных сетей для выделения и сжатия акустической информации, значимой для распознавания, а также связей для выявления временных зависимостей между фонемами в целях выработки модели слова в бурятском языке. Процент неправильно распознанных фонов составляет 11%, что существенно превышает показатели, характерные для современных моделей. Ограничением данной модели является недостаточный объём обучающих данных и зависимость от голоса диктора. Расширение банка речевых образцов, их аннотирование и оптимизация модели могут повысить точность системы и её устойчивость к вариативности речи.

Ключевые слова: бурятский язык, малоресурсные языки, речевые технологии, автоматическое распознавание речи

© Чимитов Б. Б., Завьялова В. Л. 2025

Для цитирования: Чимитов Б. Б., Завьялова В. Л. Дикторозависимое автоматическое распознавание речи на бурятском языке // Теоретическая и прикладная лингвистика. 2025. Вып. 11, № 2. С. 166–177. <https://doi.org/10.22250/24107190-2025-11-2-166>

Belig B. Chimitov✉, **Viktoriya L. Zavyalova**
Far Eastern Federal University
Vladivostok, Russian Federation
chimitov.bb@dvmfu.ru

Speaker-dependent automatic speech recognition in Buryat

Abstract

Due to the dominance of Russian in education, formal communication, and mass media the usage of Buryat is gradually decreasing. Therefore, the problem of language preservation is of high concern for the native people. In modern conditions, in resolving this question significant consideration is given to speech technologies based on natural language processing, artificial intelligence and other methods to digitize the Buryat language and provide access to resources online. The analysis of current resources has found no Buryat-targeted technologies that work with audio data. Therefore, the present study covers the development of speech recognition model prototype in

Buryat for automatic speech processing. The model uses the DeepSpeech2 architecture allowing to combine the functions of convolutional neural networks to extract and compress acoustic information relevant for recognition, as well as connections to identify time dependencies between phones in order to develop a model of a word in the Buryat language. The percentage of incorrectly recognized phones is 11%. This value significantly exceeds the values typical for modern models. The limitation of this model is the insufficient amount of training data and the dependence on the speaker's voice. Expanding the speech sample bank, annotating the data and optimizing the model can improve the accuracy of the system and its resistance to speech variability.

Keywords: Buryat language, low-resource languages, speech technologies, automatic speech recognition

© Chimitov B. B., Zavyalova V. L. 2025

For citation: Chimitov, B. B., & Zavyalova, V. L. (2025). Diktorozavisimoe avtomaticheskoe raspoznavanie rechi na buryatskom yazyke [Speaker-dependent automatic speech recognition in Buryat]. *Teoreticheskaya i prikladnaya lingvistika [Theoretical and Applied Linguistics]*, 11 (2), 166–177. <https://doi.org/10.22250/24107190-2025-11-2-166>

1. Введение [Introduction]

По данным издания ЮНЕСКО бурятский язык относится к категории уязвимых (*vulnerable*) языков [Moseley, 2010, p. 50]. Несмотря на рост численности этнического населения, количество носителей языка постепенно уменьшается. На национальных территориях Республики Бурятия, Иркутской области, Забайкальского края использование бурятского языка ограничено семейно-бытовым общением, в то время как в сферах деловой коммуникации, образования и средствах массовой информации преобладает русский язык. Таким образом, текущая ситуация характеризуется тенденцией к отказу от бурятского языка в пользу доминирующего русского языка [Дырхеева, 2018, с. 318].

В настоящее время принимаются меры по сохранению бурятского языка. В вопросе поддержания его функциональности особое внимание уделяется его диалектным формам, поскольку в сравнении с городами бурятский язык является преобладающим средством устного общения между жителями сельских регионов. По этой причине сохранение диалектных форм имеет первостепенное значение. По мнению Ж. Б. Бадагарова особую роль в восстановлении языка будут играть языковые технологии [Interview..., 2016]. Примером может послужить создание диалектного корпуса с целью сохранения звучащей речи в цифровом формате, предоставления доступа, возможности поиска, прослушивания и исследования диалектного разнообразия устной речи.

Построение подобного корпуса требует большого объема речевых образцов, подбор которых представляет относительно меньшую трудность, благодаря свободным интернет-ресурсам. Проблема заключается в необходимости составления письменных транскриптов для каждого образца речи, требующих ручного аннотирования. Альтернативным решением является использование систем автоматического распознавания речи, призванных значительно сократить ручную работу и обеспечить сбор необходимых данных в нужном формате.

Как правило, наличие большого объема аннотированных в цифровом формате речевых данных открывает возможности для разработки широкого спектра речевых технологий. Однако позднее появление бурятского языка в цифровом пространстве способствовало задержке интеграции подобных систем. На текущий момент, как и у многих языков мира, в сети Интернет в открытом доступе функционирует и ежегодно пополняется письменный корпус литературного бурятского языка, включающий 3 миллиона словоупотреблений [Бурятский корпус, 2025]. Кроме того, существует параллельный корпус бурятского языка в национальном корпусе русского языка, насчитывающий

400 тысяч слов [НКРЯ, 2025]. С 2024 года бурятский язык стал доступен в переводчике от компании Google [Google Translate, 2025], однако опции голосового ввода и прослушивания недоступны. Анализ ресурсов позволяет сделать вывод, что обученных моделей распознавания речи на бурятском языке на данный момент не существуют. Поэтому целью исследования является разработка акустической модели бурятского языка. На данном этапе рассматривается прототип модели для дикторозависимого распознавания речи.

2. К проблеме автоматического распознавания речи на бурятском языке [Automatic Buryat speech perception challenges]

Под автоматическим распознаванием речи понимают процесс преобразования устной речи в текстовую форму посредством компьютерных вычислений. На высоком уровне эта операция представляет собой функцию f , которая находит соответствие между акустическими данными X и последовательностью слов W [Kamath et al., 2019, p. 369]. Соответственно, задача распознавания сводится к поиску функции f .

Современные системы автоматического распознавания речи всё чаще основываются на методах глубокого обучения нейронных сетей. Обучение нейронных сетей позволяет аппроксимировать функцию f посредством предоставления модели входных и целевых данных. Так, распознавание речи подразумевает поиск зависимости между входными акустическими признаками и единицами распознавания.

Несмотря на то, что желаемой формой выходных данных является последовательность распознанных слов (транскрипт), в условиях ограниченного количества аннотированных данных встречаемость отдельных лексических единиц зачастую недостаточна, чтобы учесть всевозможные реализации произношения слов. Поэтому прибегают к решению дискретизировать единицу распознавания до элементов сегментного уровня языка – минимальных звуковых единиц.

2.1. Звуковой строй бурятского языка [Buryat sound system]

Фонетическая структура современного бурятского языка включает 18 гласных /*a, o, u, э, ү, и, aa, oo, уу, ээ, өө, үү, ии, ы, ай, оэ, уи, үи*/ и 27 согласных фонем /*n, n', б, б', м, м', т, т', с, ш, д, д', з, ж, н, н', р, р', л, л', й, г, г', нг, х, х', h*/ [Radnaeva, 2006, p. 578]. Классификация гласных бурятского языка различает фонемы по следующим признакам:

- 1) длительность: краткие и долгие гласные;
- 2) стабильность артикуляции: монофтонги и дифтонги;
- 3) участие губ: огубленные и неогубленные;
- 4) горизонтальное положение языка: передние, передние-продвинутые назад, задние-продвинутые вперед, задние;
- 5) вертикальное положение языка: верхние, средние, нижние [Radnaeva, 2003, p. 2738].

Основы классификации фонем бурятского языка были заложены ещё И. Д. Бураевым в 1950-е годы и, применительно к гласным звукам, классификация подразумевала выделение гласных переднего, центрального, смешанного и заднего рядов [Бураев, 1959, с. 107]. Выделение центрального и смешанного рядов было связано со средним положением языка в полости рта и тенденцией к подъёму либо одной части, либо всей спинки языка [Бураев, 1959, с. 111]. Тем не менее, все среднеязычные фонемы относились к центральному ряду, в то время как смешанный ряд включал лишь некоторые аллофонные оттенки звуков других рядов. С другой стороны, выделение переднего-отодвинутого и заднего-продвинутого рядов позволяет разграничивать такие гласные как /*э, ээ*/ и /*и, ии*/, а также фонему /*ы*/, которая ранее в классификации И. Д. Бураева не выделялась и считалась оттенковым звуком.

Система согласных фонем бурятского языка различает звуки по признакам:

- 1) по способу образования: взрывные, фрикативные, носовые и дрожание сонанты;
- 2) по месту образования: губные, альвеолярные, среднеязычные-палатальные, велярные и гортанные;
- 3) по звонкости: глухие и звонкие;
- 4) по мягкости: мягкие и твёрдые;
- 5) по источнику звука: шумные и сонанты [Radnaeva, 2003, p. 2739].

Наличие гортанной фонемы /h/ является отличительной особенностью бурятского языка от собственно монгольского языка и других представителей монголоязычного семейства. Также в системе согласных фонем бурятского языка отсутствуют аффрикаты. Данные признаки наряду с другими особенностями обеспечили ответвление и самостоятельный статус бурятского языка [Бураев, 2004, с. 229].

Проблема использования фонем как единиц распознавания заключается в неоднородности их реализаций в потоке речи. В условиях естественной речи зачастую реализуются аллофоны, отличные от эталонов, определяемых классическими правилами реализации звуков языка в потоке речи. Отчасти эта проблема может решаться путём фонетического транскрибирования, при котором учитываются особенности реализации фонем в зависимости от конкретных фонетических условий, например, в ударных и безударных позициях, учитываются влияния соседних звуков, в особенности предшествующих согласных звуков [Попов, 2014, с. 19].

Применительно к гласным аллофоны той или иной фонемы могут встречаться в зависимости от расположения в первом и непервом слогах [Radnaeva, 2003, p. 2738]. Одними из основных закономерностей, выделяемых Л. Д. Раднаевой, которые применяются к гласным звукам в естественной речи, являются:

- 1) «длительность гласных (кратких, долгих и дифтонгов) в первом слоге превышает длительность гласных в других слогах, кроме последнего;
- 2) длительность долгих гласных и дифтонгов всегда превышает длительность кратких гласных в любой позиции слова;
- 3) длительность кратких гласных в середине слова всегда короче длительности краткого первого и последнего слога; и др.» [Radnaeva, 2006, p. 579].

Учитывая правила, можно сделать вывод о характере количественной редукации гласных, кроме долгих и дифтонгов, внутри непервого и непоследнего слогов. Зачастую относительно длительная артикуляция гласных в начале и конце, а также долгих гласных и дифтонгов в срединных слогах, приводит к ощущению наличия в рамках одного слова двойного и тройного ударения [Radnaeva, 2006, p. 580].

Во-вторых, аллофоны гласных проявляются в зависимости от соседних звуков. Например, после альвеолярных смычных согласных артикуляция гласных производится несколько ближе к передней части ротовой полости. Напротив, после заднеязычных велярных согласных артикуляция смещается назад. Под влиянием губных согласных гласные, как правило, огубляются [Radnaeva, 2003, p. 2739].

В-третьих, оттенковые признаки согласных звуков также проявляются в зависимости от начального, конечного и интервокального положения [Radnaeva, 2003, p. 2739]. Как и в русском языке, характерной особенностью бурятского языка является оглушение конечных звонких звуков, а также в позиции перед глухими согласными.

Губные фонемы /b/, /b'/ и велярный /z/ могут иметь оттенковый признак щелинности. Данная особенность наблюдается в интервокальной позиции, а также после сонорных /p/ и /l/. Например, в словах *аба* (отец), *абяан* (звук), *угэ* (слово) вместо смычных /b/ и /b'/ реализуются звуки близкие к /v/ и /v'/ русского языка, а вместо смычного /г/ реализуется щелевой аллофон [Бураев, 1959, с. 62].

Несмотря на большой охват возможных звуковых реализаций, фонемы вместе с аллофонами являются звуковыми единицами фонетической системы, что связывает их в одном конкретном языке. Напротив, в качестве «абсолютных» единиц, выделяемых в речевом потоке, используют понятие фонов. В отличие от фонем и аллофонов, которые являются языковыми единицами, фон представляет собой индивидуальный звук речи, конкретную реализацию фонемы. В отличие от фонем, фонны не привязаны к какому-либо конкретному языку, а являются строительными блоками самой речи [Kamath et al., 2019, p. 378].

Применительно к данному исследованию аннотирование данных на уровне фоннов позволяет выделить сегменты, которые будут учитывать оттенковые признаки гласных и согласных звуков бурятского языка, а также будут охватывать звуковые сегменты, не принадлежащие к бурятскому языку. Например, в исконно бурятских словах отсутствуют звуки /ц, ч, щ/, что связано с отсутствием аффрикат в звуковом строе. Однако их немало в словах, заимствованных из русского языка.

С другой стороны, использование фоннов в последующем позволит выполнять несколько вариантов аннотации для одного и того же слова при конвертации транскрипции в результирующий текст. Предполагается, что такой подход обеспечит устойчивость модели к диалектным вариантам произношения одних и тех же единиц.

Таким образом, в данном исследовании мы дискретизируем единицу распознавания до уровня фоннов. Для текстовой репрезентации мы применяем нотацию CMU Sphinx [Kamath et al., 2019, p. 378]. Поскольку данная нотация изначально была ориентирована на английский язык, применительно к бурятскому языку состав фоннов был дополнен. Список пар фонем-фоннов представлен ниже.

Гласные:

/a/ – А; /aa/ – АА; /o/ – О; /oo/, /θθ/ – ОО; /y/, /ỵ/ – У; /yy/, /ỵỵ/ – УУ; /э/ – Е; /ээ/ – ЕЕ; /u/ – I; /uu/ – II; /ы/ – Y; редуц. гласная – EH.

Согласные:

/b/ – В; /ḅ/ – ВJ; /β/ – V; /β̣/ – VJ; /z/ – G; /ẓ/ – GJ; фриктивный /z/ – Gr; /d/ – D; /ḍ/ – DJ; /ʒ/ – ZH; /z/ – Z; /ẓ/ – ZJ; /j/ – J; /k/ – K; /ḳ/ – KJ; /l/ – L; /ḷ/ – LJ; /m/ – M; /ṃ/ – MJ; /n/ – N; /ṇ/ – NJ; /n̄/ – NG; /n/ – P; /ṇ/ – PJ; /p/ – R; /p̣/ – RJ; /c/ – S; /c̣/ – SJ; /m/ – T; /ṃ/ – TJ; /f/ – F; /f̣/ – FJ; /x/ – H; /x̣/ – HJ; /h/ – HH; /ɥ/ – C; /ɥ/ – CH; /u/ – SH; /ụ/ – SCH

В данном списке фонемы /y/, /ỵ/ ставятся в соответствие одному фонону U. Несмотря на то, что звуки различаются по месту артикуляции, качественно они являются похожими, а в речи зачастую происходит их уподобление под влиянием соседних переднеязычных согласных: гласная заднего ряда /y/ по позиции приближается к артикуляции гласной заднего-продвинутого ряда /ỵ/ [Бураев, 1959, с. 176]. Аналогично фонемы /oo/, /θθ/ ставятся в соответствие с фоном O. Также в списке указаны звуки, принадлежащие к русскому языку – /в, в', ф, ф', ц, ч, щ/.

2.2. Существующие методики предобработки данных [Existing data pre-processing techniques]

Обработка речевых данных предполагает преобразование звука речи в цифровой вид, причём конечная репрезентация должна удовлетворять требованиям как сжатости, так и ёмкости кодируемой информации. Кодирование звука заключается в преобразовании колебаний воздуха в колебания электрического тока с последующей операцией дискретизации полученного аналогового электрического сигнала. Выходом преобразования является одномерный массив чисел, которые представляют собой оцифрованный звуковой сигнал, определённый во временной области. Данный вид позволяет хранить звукозапись в электронном виде, но ввиду своей многомерности представляет сложность при

моделировании. Поэтому в речевых системах сигнал проходит дополнительную обработку, и чаще подобные операции основываются на спектральном анализе [Kamath et al., 2019, p. 371].

Спектральный анализ позволяет выделить частотные составляющие внутри акустического сигнала. Преобразование из временной в частотно-временную область производится с помощью оконных преобразований Фурье (Short-Time Fourier Transform). Результатом данной операции является спектрограмма, позволяющая проследить изменения частотных компонентов в зависимости от времени [Kamath et al., 2019, p. 374].

Спектрограмма является наиболее ёмкой частотно-временной репрезентацией акустического сигнала. Одним из способов сжатия спектральной информации является использование фильтров Мела, по итогу на выходе получается Мел-спектрограмма. В отличие от классической спектрограммы, которая учитывает логарифмическое восприятие интенсивности звука слуховой системой человека, Мел-спектрограммы поверх этого имитируют высокую дифференциацию человеческим ухом низкочастотных сигналов по сравнению с высокочастотными составляющими. В связи с этим, вводится шкала Мела, которая устанавливает соответствие между физической величиной измерения частоты (герц) и психологической величиной измерения частоты (мел) [Stevens et al., 1937]. С другой стороны, в сравнении с классической спектрограммой, разрешение которой по оси частот обычно составляет 1025 единиц (частотных ячеек, frequency bins), разрешение Мел-спектрограмм по той же оси обычно не превышает 256 единиц, тем самым достигается многократное сжатие исходного сигнала.

2.3. Существующие практики разработки акустических моделей с использованием нейронных сетей [Designing speech acoustic models with the use of neural nets]

Среди методов обработки речевых сигналов особое место занимают методы с использованием нейронных сетей [Алимурадов, 2015, с. 31]. Будучи частью сферы машинного обучения, подобные технологии способны самостоятельно выявлять закономерности внутри тренировочных данных без необходимости явного указания инструкций по решению комплексных задач, таких как классификация, регрессия и кластеризация. Применительно к задачам распознавания речи и изображений, к обработке естественной речи были разработаны более сложные типы нейронных сетей, использующие более продвинутые вычисления с многомерными данными, что привело к развитию методов глубокого обучения [Janiesch et al., 2021, p. 687].

Среди доступных архитектур акустических моделей перечисляют DeepSpeech2 [Amodei et al., 2016], Wav2Vec [Baeovski et al., 2020], Wav2Letter [Collobert et al., 2016] и другие, являющиеся передовыми моделями в сфере распознавания речи. Изначально обученные для распознавания английского языка модели в последующем были успешно адаптированы для распознавания других языков мира, в том числе и русского языка.

Несмотря на существование более современных альтернативных вариантов, достоинством архитектуры DeepSpeech2 является сочетание как свёрточных, так и рекуррентных нейронных сетей [Amodei et al., 2016]:

1) Свёрточные нейронные сети (Convolutional neural networks CNN) – широко применяются в распознавании изображений и речи, благодаря своей особенности сжимать и выявлять существенные признаки внутри входного сигнала.

2) Рекуррентные нейронные сети (Recurrent neural networks RNN) – спроектированы для анализа динамических последовательностей данных, в которых важное значение играют временные зависимости, тем самым предоставляется возможность учитывать контекст [Janiesch et al., 2020, p. 690].

В данном исследовании для разработки прототипа акустической модели бурятского языка используется DeepSpeech2-подобная архитектура. Среди достоинств модели выделяются её стойкость к незначительным вариациям произношения путём фильтрации входной информации за счёт операции свёртки, а также возможность учёта фонетического контекста при распознавании речевых единиц в определённый момент времени.

2.4. Метрики оценки качества модели распознавания речи **[Quality assessment metrics for speech recognition model]**

Оценка качества распознавания производится путём подсчёта стандартных показателей, основным из которых является процент неправильно распознанных слов (WER – Word Error Rate). В случаях посимвольного распознавания может также применяться процент неправильно распознанных символов (CER – Char Error Rate). Таким образом, целью построения моделей является уменьшение показателя WER в сторону нуля.

Метод определения WER состоит в подсчёте количества операций замены, удаления и вставки для редактирования одной текстовой строки (результата распознавания) в другую (истинной последовательности слов) путём вычисления расстояния Левенштейна на уровне слов. Конечная формула расчёта показателя имеет следующий вид:

$$WER = \frac{S+D+I}{T} \cdot 100\%$$

, где S – количество операций замены, D – количество удалений, I – количество вставок, T – количества слов в распознаваемой фразе [Тампель, 2016, с. 43].

3. Материал и методика исследования [Material and methods]

3.1. Материал [Material]

В соответствии с поставленной целью исследования по разработке дикторозависимой системы были собраны образцы речи одного диктора-носителя для обучения прототипа акустической модели бурятского языка. Диктор является русско-бурятским билингом, носителем хоринского диалекта, агинского говора. Возраст диктора составляет 24 года, пол – мужской.

Диктору были предложены для чтения новостные статьи из газетной редакции «Толон» [Толон, 2025] и новостного сайта «Соёл» [Соёл, 2025]. Аудиозапись осуществлялась в нестудийных условиях на записывающее устройство. Средняя продолжительность одного образца составляет 6 секунд, общая продолжительность данных – 2.5 часа (1500 + 30 образцов).

3.2. Методика [Methods]

Ход работы включал выполнение несколько последовательных этапов.

3.2.1. Предобработка звуковых данных [Speech data pre-processing]

Первым шагом была проведена предобработка речевых данных. Для каждого образца речи была вычислена соответствующая Мел-спектрограмма. При расчёте были использованы следующие параметры:

- 1) частота дискретизации (sample rate) = 16 кГц;
- 2) количество Мел-фильтров (mel filter banks) = 80;
- 3) размер окна (window size) = 25 мс;
- 4) размер перекрытия (hop length) = 10 мс.

Вычисление спектрограмм производилось с помощью библиотеки torchaudio [Torchaudio..., 2025], предназначенной для обработки звуковых данных. Библиотека предоставляет инструменты в том числе для вычисления Мел-спектрограмм с указанием вышеперечисленных параметров. С помощью этой же библиотеки численные значения полученных Мел-спектрограмм были далее прологарифмированы в целях получения значений в децибелах.

Конечным действием предобработки данных была произведена нормализация и стандартизация данных. Поскольку каждая Мел-спектрограмма представляет собой массив чисел, большая разбросанность допустимых значений может значительно увеличить время обучения. Поэтому значения нормализуются: среднее значение данных равно 0, среднеквадратичное отклонение – 1 [Lecun et al., 2000, p. 16].

3.2.2. Обучение разрабатываемой модели распознавания речи [Recognition model training]

Для разработки прототипа модели распознавания была использована DeepSpeech2-подобная архитектура нейронной сети. Общее количество параметров составляет 377 тысяч, что представляет лишь малую долю от оригинальной модели DeepSpeech2 (38 миллионов параметров). Модель обучалась на 1500 образцах в течение 1 часа.

3.2.3. Тестирование и определение точности модели [Testing for the model accuracy]

В тестовых испытаниях были использованы 30 образцов речи того же диктора, которые не применялись на стадии обучения. Результаты оценивались по метрикам Word Error Rate с учётом того, что сравнивались последовательности распознанных и аннотированных фонов и, соответственно, вычислялся процент неправильно распознанных фонов.

4. Обсуждение результатов [Results and discussion]

По результатам тестирования на 30 образцах речи вычисление по метрике WER процента неправильно распознанных фонов составило 11% – в среднем каждый десятый фон должен был быть заменён или удалён для соответствия целевой последовательности фонов. Также в данный показатель входит количество недораспознанных (пропущенных) фонов. Значение заметно превышает стандартные показатели метрики WER. Например, дикторонезависимая модель распознавания английской речи Wav2Vec для английского языка имеет показатель 3.3%.

Примеры неточного распознавания моделью приведены ниже (рис. 1), где символ «|» указывает на границы между отдельными словами (символ пробела), а символ «-» указывает на пропуск значимого фона (недораспознавание).

Во-первых, модель ошибочно распознаёт единицы, которые отсутствуют в исходном сигнале: слово *аглаг*, /A G L EH K/, распознаётся с присутствующим фарингальным звуком /h/ в начале как *hаглаг* /HH A G L EH K/.

Во-вторых, модель пропускает некоторые значимые фоны, как, например, в слове *нютагтаа* /NJ U T EH K T AA/, пропущена часть корня слова нютаг (Родина).

В-третьих, из-за неверной постановки разделителей (рис. 2), происходит наслаивание слов между собой: *аглаг тэнюун* распознаётся как *аглагтэ нюун*.

В-четвёртых, в интервокальных позициях модель не различает звуки по признаку звонкости: *тубхинэгты* (перевод: *поселитесь*) /T U P HJ I N EH K T Y/ распознаётся как /D U P HJ I N EH K T Y/.

Аглаг тэнюун нютагтаа айл боложо түбхинэгты

A G L E H K | T E N J U U N G | N J U T E H K T A A | E E L | B O L E H Z H E H | T U P H J I N E H K T Y

НН A G L E H K T E | N J U U N G - - - T E H K T E H | - L | B O L E H Z H E H | D U P H J I N E H K T Y

Р и с у н о к 1. Пример распознавания [F i g u r e 1. Recognition sample result]

Примечание к рисунку 1: Исходное предложение (сверху), аннотированная последовательность фонем (посередине), распознанная последовательность фонем (снизу)

Амжалтатай, урагшатай ябаг лэ!

A M Z H E H L T E H T E E | U R E H G S H E H T E E | J A V E H G | L E H

A M Z H E H L T E H | T E E | U R E H G S H E H T E E | J A V E H G L E H

Р и с у н о к 2. Пример некорректной постановки границ между словами [F i g u r e 2. Erroneous segmentation of words]

Примечание к рисунку 2: Исходное предложение (сверху), аннотированная последовательность фонем (посередине), распознанная последовательность фонем (снизу)

На примере предложения *Аглаг тэнюун нютагтаа айл боложо түбхинэгты (Постройте семью на своей спокойной койной Родине – перевод Б. Б. Чимитова)* можно сделать несколько выводов относительно точности модели (рис. 1–2). Корректировка неверно распознанных границ слов представляет собой сложную задачу, поскольку символ пробела сам по себе не имеет чёткого акустического признака в отличие от звуков, которые на спектрограмме однозначно дифференцируются формантами F1, F2, формантными переходами, своеобразным шумом и т. д. Отчасти отсутствие сигнала (тишины) может указывать на границу слова, однако естественная речь характеризуется слитностью, при которой паузы наблюдаются только между цельными фразами. Поэтому модель помимо установления соответствия между акустическими признаками и звуковыми единицами, дополнительно вырабатывает модель слова за счёт рекуррентных связей внутри своей архитектуры. Из-за ограниченного количества уникальных слов внутри обучающих данных есть большая вероятность некорректной постановки разделительного знака при распознавании не встречающихся ранее слов. На данный момент количество уникальных словоформ составляет 9000 единиц.

5. Заключение [Conclusion]

Современные модели распознавания речи позволяют успешно преобразовывать устную речь в текстовый формат. Подобные речевые технологии могут быть использованы для эффективного сбора данных малоресурсных языков, к которым относится бурятский язык. Учитывая проблему исчезновения языка, и важную роль в решении данного вопроса диалектов бурятского языка возникает большая потребность в использовании подобных моделей распознавания речи для автоматического аннотирования живой акцентной речи. В последующем данные послужат необходимым материалом для создания диалектного корпуса с целью сохранения звучащей речи, предоставления доступа, возможности поиска, прослушивания и исследования диалектного разнообразия устной речи бурятского языка. Кроме того, наличие данных в цифровом формате будет способствовать разработке и развитию широкого спектра речевых технологий, которые на данный момент доступны только доминирующим языкам.

В работе были рассмотрены разработка и обучение прототипа модели распознавания речи на бурятском языке на речевых данных одного диктора. Анализ существующих практик машинного обучения в сфере автоматического распознавания речи позволяет сделать несколько выводов, которые будут учитываться в дальнейшем для построения дикторонезависимой системы. Во-первых, дискретизация единиц распознавания до минимальных звуковых единиц языка позволяет обучать акустические модели в условиях ограниченности словоформ как цельных единиц внутри тренировочных данных. Посредством разбиения на звуковые единицы возрастает количество обучающих примеров реализаций звуков в разных фонетических контекстах. Напротив, одновременно ставится требование вырабатывания системой модели слова, за счёт чего определяется принадлежность последовательности звуковых единиц к одному слову.

В данном исследовании дискретизация единиц распознавания базируется на звуках речи – фонах. Использование фонов вместо фонем бурятского языка в качестве распознаваемых единиц позволяет учитывать оттенковые реализации согласных и гласных фонем в разных фонетических контекстах. Кроме того, используемый набор фонов предусматривает звуковые единицы /в, в', ф, ф', ц, ч, ц'/, исконно принадлежащие русскому языку и встречающиеся только в заимствованных словах. Общее количество распознаваемых на данный момент фонов составляет 52: 12 гласных и 39 согласных, а также разделительный символ (символ пробела). В зависимости от используемой нотации, а также оптимизационных решений данное количество может корректироваться.

Во-вторых, выбор репрезентации звуковых данных является одним из основных факторов, влияющих на количество необходимых ресурсов памяти и время обучения модели. В качестве репрезентаций были использованы Мел-спектрограммы, которые определяют акустические признаки в частотно-временной области, а также являются сжатой формой кодирования звуковой информации по сравнению с классическими спектрограммами. Кроме того, Мел-спектрограммы позволяют учитывать не только логарифмическое восприятие громкости, но и нелинейное восприятие тона звука слуховой системой человека.

Разработанная дикторозависимая модель распознавания бурятской речи может послужить прототипом для разработки дикторонезависимой системы. Используемая в модели архитектура DeepSpeech2 комбинирует особенности свёрточных нейронных сетей по выделению и сжатию значимой для распознавания акустической информации, а также рекуррентных связей по выявлению временных зависимостей между фонами в целях выработки модели слова в бурятском языке. По мере увеличения данных параметры могут также корректироваться для обеспечения точности и устойчивости модели.

Вычисленное значение процента неправильно распознанных фонов составляет 11%, что значительно выше, чем у широко используемых современных моделей. Предполагается, что увеличение параметров модели и увеличение обучающих данных будут способствовать уменьшению количества ошибок.

Перспектива дальнейшей работы заключается в расширении банка речевых данных, их аннотирования и оптимизации модели. Увеличение количества уникальных голосов будет способствовать устойчивости модели к вариациям голоса. В целях повышения точности системы и конвертации выводов модели в результирующий транскрипт дополнительно планируется внедрение алгоритмов декодирования и разработки языковой модели.

Библиографический список

- Бураев, 1959 – Бураев И. Д. Звуковой состав бурятского языка. Улан-Удэ : Изд-во Бурятского комплексного науч.-исслед. ин-та, 1959. 193 с.
- Бураев, Шагдаров, 2004 – Бураев И. Д., Шагдаров Л. Д. Образование бурятского языка // Буряты / отв. ред. Л. Л. Абаева, Н. Л. Жуковская. М. : Наука, 2004. С. 227–231.

- Бурятский корпус, 2025 – Бурятский корпус: письменный корпус бурятского языка. 2025. URL : <https://buriyat.web-corpora.net> (дата обращения : 03.02.2025).
- Дырхеева, 2018 – Дырхеева Г. А. Языковая ситуация в Республике Бурятия // Мир Большого Алтая. 2018. № 4 (2). С. 302–320.
- НКРЯ, 2025 – Национальный корпус русского языка. 2025. URL : <https://ruscorpora.ru/> (дата обращения : 03.02.2025).
- Попов, 2014 – Попов М. Б. Фонетика современного русского языка : учебник. СПб. : Филол. факультет СПбГУ, 2014. 303 с.
- Соёл, 2025 – Соёл : новостной сайт. 2025. URL : <https://soyol.ru/bur/> (дата обращения : 03.02.2025).
- Тампель, Карпов, 2016 – Тампель И. Б., Карпов А. А. Автоматическое распознавание речи : учеб. пособие. СПб. : Ун-т ИТМО, 2016. 138 с.
- Толон, 2025 – Редакция газеты «Толон» в социальных сетях. 2025. URL : <https://vk.com/tolon07> (дата обращения: 03.02.2025).
- Amodei et al., 2016 – Amodei D., Ananthanarayanan S., Anubhai R., Bai J., Battenberg E., Bell J., Case C., Casper J., Catanzaro B., Cheng Q. Deep speech 2: End-to-end speech recognition in English and Mandarin // Proc. of the 33rd International Conference on Machine Learning (ICML). New York, 2016. Vol. 2. P. 173–182.
- Baevski et al., 2020 – Baevski A., Zhou H., 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.
- Collobert et al., – Collobert R., Puhersch C., Synaev G. Wav2Letter: An end-to-end ConvNet-based speech recognition system // Computing research repository. URL: <https://arxiv.org/pdf/1609.03193> (дата обращения : 03.02.2025).
- Google translate, 2025 – Google translate: Buryat–Russian. 2025. URL : <https://translate.google.com/?hl=ru&sl=bur&tl=ru&op=translate> (дата обращения : 03.02.2025).
- Interview with Jargal Badagarov, 2016 – Interview with Jargal Badagarov – leader of a new Thematic Network. 2016. URL : <https://www.uarctic.org/news/2016/12/interview-with-jargal-badagarov-leader-of-a-new-thematic-network> (дата обращения : 03.02.2025).
- Janiesch et al., 2021 – Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning // Electron Markets. 2021. Vol. 31. P. 685–695.
- Kamath et al., 2019 – Kamath U., Liu J., Whitaker J. Deep Learning for NLP and speech recognition. New York : Springer Publishing Company, Incorporated, 2019. 621 p.
- LeCun et al., 2000 – Lecun Y., Bottou L., Orr G. B., Müller K. Neural Networks : Tricks of the Trade. Efficient BackProp. Berlin : Springer, 2000. 769 p.
- Moseley, 2010 – Atlas of the World’s languages in danger, 2010. URL : <https://unesdoc.unesco.org/ark:/48223/pf0000187026> (дата обращения : 03.02.2025).
- Radnaeva, 2006 – Radnaeva L. D. Phonetic research of the sound form of modern Buryat language // SPECOM. St Petersburg, 2006. P. 577–582.
- Radnaeva, 2003 – Radnaeva L. D. Two types of transcription (The Buryat Language Database) // ICPHS. Barcelona, 2003. P. 2737–2740.
- Stevens et al., 1937 – Stevens S. S. Volkman J., Newman E. B. A scale for the measurement of the psychological magnitude pitch // The Journal of the Acoustical Society of America. 1937. Vol. 8. P. 185–190.
- TorchAudio, 2025 – TorchAudio Documentation. URL: <https://pytorch.org/audio/stable/index.html> (дата обращения : 03.02.2025).

References

- Buraev, I. D. (1959). *Zvukovoy sostav buryatskogo yazyka [Sound system of the Buryat language]*. Ulan-Ude : Buryat Complex Research Institute Press. (In Russ.).
- Buraev, I. D., & Shagdarov, L. D. (2004). *Obrazovanie buryatskogo yazyka [Development of the Buryat language]*. In L. L. Abaeva, & N. L. Zhukovskaya (Eds), *Buryaty [The Buryat]* (pp. 227–231). Moscow : Nauka Press. (In Russ.).

- Buryat corpora. (2025). *Buryatskiy korpus: pis'mennyy korpus buryatskogo yazyka [Buryat corpora: written corpora of the Buryat language]*. Retrieved February 3, 2025 from <<https://buryat.web-corpora.net/>>.
- Dyrkheeva, G. A. (2018). Yazykovaya situatsiya v Respublike Buryatiya [The language situation in Buryat Republic]. *Mir Bol'shogo Altaya [World of the Great Altai]*, 4, 302–320. (In Russ.).
- Russian Corpus. (2025). *Natsional'nyy korpus russkogo yazyka [Russian National Corpus]*. Retrieved February 3, 2025 from <<https://ruscorpora.ru/>>.
- Popov, M. B. (2014). *Fonetika sovremennogo russkogo yazyka [Phonetics of the modern Russian language]: A coursebook*. Philological Faculty of St Petersburg State University Press. (In Russ.).
- Soyol. (2025). *Novostnoy sayt «Soyol» [News website «Soyol»]*. Retrieved February 3, 2025 from <<https://soyol.ru/bur/>>.
- Tampel, I. B., & Karpov, A. A. (2016). *Avtomaticeskoe raspoznavanie rechi [Automatic speech recognition]: A coursebook*. St Petersburg ITMO University Press. (In Russ.).
- Tolon. (2025). *Redaktsiya gazety «Tolon» v social'nykh setyakh [Tolon Press in social networks]*. Retrieved February 3, 2025 from <<https://vk.com/tolon07>>.
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Bell, J., Case, C., Casper, J., Catanzaro, B., & Cheng, Q. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. *Proc. of the 33rd International Conference on Machine Learning (ICML)* (Vol. 2, pp. 173–182). New York.
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.
- Collobert, R., Puhersch, C., & Synaeva, G. (2016). Wav2Letter: an End-to-End ConvNet-based Speech Recognition System. *Computing research repository*. Retrieved February 3, 2025 from <<https://arxiv.org/pdf/1609.03193>>.
- Google translate (2025). *Google translate: Buryat–Russian*. Retrieved February 3, 2025 from <<https://translate.google.com/?hl=ru&sl=bua&tl=ru&op=translate>>.
- Interview with Jargal Badagarov. (2016). Interview with Jargal Badagarov – leader of a new Thematic Network. Retrieved February 3, 2025 from <<https://www.uarctic.org/news/2016/12/interview-with-jargal-badagarov-leader-of-a-new-thematic-network>>.
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electron Markets*, 31, 685–695.
- Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep Learning for NLP and Speech Recognition*. New York : Springer Publishing Company, Incorporated.
- Lecun, Y., Bottou, L., Orr, G. B., & Müller, K. (2000). *Neural Networks: Tricks of the Trade. Efficient BackProp*. Berlin : Springer.
- Moseley, C. (2010). Atlas of the World's languages in danger. Retrieved February 3, 2025 from <<https://unesdoc.unesco.org/ark:/48223/pf0000187026>>.
- Radnaeva, L. D. (2006). Phonetic Research of the Sound Form of Modern Buryat Language. *SPECOM* (pp. 577–582). St Petersburg.
- Radnaeva, L. D. (2003). Two types of transcription (The Buryat Language Database). *ICPhS* (pp. 2737–2740). Barcelona.
- Stevens, S. S. Volkman, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8, 185–190.
- TorchAudio Documentation. (2025). Retrieved February 3, 2025 from <<https://pytorch.org/audio/stable/index.html>>.