НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2026 Том 26 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2026 Vol. 26 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-4-826-834
УДК 004.522

Система распознавания карельской речи с переключением кода карельский-русский

Ирина Сергеевна Кипяткова[1]✉, Михаил Дмитриевич Долгушин[2], Ксения Олеговна Киселева[3], Ильдар Амирович Кагиров[4]

[1,2,3,4] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] kipyatkova@iias.spb.su✉, https://orcid.org/0000-0002-1264-4458
[2] dolgushin.m@iias.spb.su, https://orcid.org/0000-0002-4344-2330
[3] kiseleva.k@iias.spb.su, https://orcid.org/0009-0001-5572-3842
[4] kagirov@iias.spb.su, https://orcid.org/0000-0003-1196-1117

Аннотация

Введение. Представлена система автоматического распознавания устной речи для ливвиковского наречия карельского языка в условиях переключения кодов между карельским и русским языками. Выполнено исследование методов автоматического распознавания билингвальной речи. С целью повышения точности распознавания разработана методика аугментации текстовых данных, включающая в себя частичный перевод и синтез словоформ с внутрисловным переключением кода. **Метод.** Применено акустическое моделирование путем дообучения предобученной многоязычной модели Wav2Vec2-BERT 2.0 на материале двух предварительно собранных баз данных общим объемом 7,5 ч. Дообучение осуществлялось с использованием фреймворка Transformers. При построении языковой модели для решения проблемы нехватки текстовых данных с переключением кода был применен метод аугментации, основанный на частичном автоматическом переводе карельских текстов на русский язык с последующей генерацией словоформ с внутрисловным переключением кода по лингвистическим правилам. На основе разработанных правил сгенерирован список слов с внутрисловным переключением кода для использования при создании модели языка. В ходе экспериментов сравнивались различные конфигурации языковой модели, различающиеся количеством итераций текстовой аугментации (от 1 до 5), содержанием словаря (базовый и включающий все сгенерированные формы), а также интерполяцией с русской языковой моделью. **Основные результаты.** Проведенные эксперименты показали, что использование полного словаря, включающего сгенерированные гибридные словоформы, дает устойчивое улучшение результатов. Дополнительное снижение количества неправильно распознанных слов до 25,82 % на отладочной и 29 % на тестовой частях базы данных было достигнуто при линейной интерполяции карельской языковой модели с русской (коэффициент интерполяции 0,7). **Обсуждение.** Выполненные эксперименты подтверждают эффективность разработанной методики создания системы распознавания билингвальной речи. В частности, рекомендуется комбинировать дообучение многоязычных акустических моделей, текстовую аугментацию с морфологическими правилами и интерполяцию языковых моделей. Предложенный подход может быть применен для создания систем распознавания речи для других малоресурсных языков России в условиях несбалансированного двуязычия.

Ключевые слова

автоматическое распознавание речи, переключение кодов, карельский язык, малоресурсные языки, аугментация данных, внутрисловное переключение кода, Wav2Vec2-BERT 2.0, языковое моделирование

Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда, проект № 24-21-00276, https://rscf.ru/project/24-21-00276/.

Ссылка для цитирования: Кипяткова И.С., Долгушин М.Д., Киселева К.О., Кагиров И.А. Система распознавания карельской речи с переключением кода карельский-русский // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 4. С. 826–834. doi: 10.17586/2226-1494-2026-26-4-826-834

Karelian speech recognition system with support for Karelian-Russian code-switching

Irina S. Kipyatkova[1]✉, Mikhail D. Dolgushin[2], Kseniia O. Kiseleva[3], Ildar A. Kagirov[4]

[1,2,3,4] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation

[1] kipyatkova@iias.spb.su✉, https://orcid.org/0000-0002-1264-4458
[2] dolgushin.m@iias.spb.su, https://orcid.org/0000-0002-4344-2330
[3] kiseleva.k@iias.spb.su, https://orcid.org/0009-0001-5572-3842
[4] kagirov@iias.spb.su, https://orcid.org/0000-0003-1196-1117

Abstract

This paper focuses on the development of an automatic speech recognition system for the Livvi-Karelian variety of the Karelian language, as it is spoken under conditions of code-switching between Karelian and Russian. The study of bilingual speech recognition methods is carried out. In order to improve the quality of speech recognition, a methodology for training text data augmentation via partial translation and intra-word code-switched wordforms artificial synthesis was developed. Acoustic modeling was performed by fine-tuning a pre-trained multilingual Wav2Vec2-BERT 2.0 model with the use of the data from two previously collected corpora containing 7.5 hours of speech. Fine-tuning was performed using the Transformers framework. When developing the language model, in order to address the problem of limited code-switching data, an augmentation method was applied based on partial automatic translation of Karelian texts into Russian, followed by the generation of word-forms with intra-word code-switching based on special linguistic rules. On the base of formulated rules, a list of words with intra-word code-switching was generated for a language model. The experiments showed that using a full vocabulary that includes generated hybrid word forms yields a consistent improvement in results. A further reduction in word error rate to 25.82 % on the development set and 29 % on the test portion of the corpus was achieved through linear interpolation of the Karelian language model with the Russian language model (interpolation weight 0.7). The conducted experiments confirm the effectiveness of the developed methodology for developing a bilingual speech recognition system. In particular, it is recommended to combine fine-tuning of multilingual acoustic models, text augmentation with morphological rules, and language model interpolation. The proposed approach can be applied to developing speech recognition systems for other low-resource languages of Russia spoken in an unbalanced bilingual environment.

Keywords

automatic speech recognition, code-switching, Karelian language, low-resource languages, data augmentation, intra-word code-switching, Wav2Vec2-BERT 2.0, language modeling

Acknowledgements

This research is financially supported by the Russian Science Foundation, project No. 24-21-00276, https://rscf.ru/project/24-21-00276/.

For citation: Kipyatkova I.S., Dolgushin M.D., Kiseleva K.O., Kagirov I.A. Karelian speech recognition system with support for Karelian-Russian code-switching. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 4, pp. 826–834 (in Russian). doi: 10.17586/2226-1494-2026-26-4-826-834

Введение

Создание систем автоматического распознавания речи для малоресурсных языков сопряжено с рядом трудностей, среди которых основными являются дефицит размеченных речевых данных и отсутствие специализированных корпусов. Задача дополнительно осложняется в тех случаях, когда объектом моделирования выступает речь с переключением кодов. Явление переключения кодов широко известно и распространено в условиях естественного полиязычия но, тем не менее, является сложным для моделирования, поскольку границы фрагментов на разных языках могут проходить не только между высказываниями, но и между словами и даже внутри одного слова.

В настоящее время на территории России насчитывается свыше 150 языков (точные цифры разнятся в зависимости от критериев включения в число языков России), при этом абсолютное большинство из них относятся к малоресурсным. Хорошим примером малоресурсного языка с переключением кодов является современный карельский. Так, в коммуникации носители карельского языка регулярно переходят на русский язык — от полных высказываний до основ слов или аффиксов. Несмотря на наличие лингвистических работ, описывающих механизм и закономерности переключения для языковой пары карельский-русский, вопросы разработки систем распознавания для карельского (и, шире, для малоресурсных языков России), остаются малоисследованными.

Целью работы является повышение качества распознавания речи на ливвиковском наречии карельского языка с переключением на русский язык за счет разработки методики аугментации обучающих текстовых данных путем частичного перевода и искусственного синтезирования словоформ с внутрисловным переключением кода.

Обзор основных методов моделирования переключения кодов в системах распознавания речи

Можно выделить два основных подхода к построению систем распознавания речи с переключением кодов [1]. Первый заключается в том, что вначале определяются границы разноязычных фрагментов речи и

соответствующие им языки, затем происходит обработка конкретного фрагмента подходящей моноязычной системой распознавания речи. Точность работы подобной системы будет зависеть от точности определения момента смены языка и точности работы блока идентификации языка [2]. В связи с этим в ходе настоящего исследования был применен второй подход, который состоит в использовании многоязычной системы распознавания речи. В этом случае акустические и языковые модели строятся сразу для обоих языков без предварительного определения языка перед процессом распознавания [3].

Для обучения моделей необходим многоязычный корпус, однако его сбор сопряжен с проблемой нехватки данных, поскольку в письменной речи переключение кодов встречается гораздо реже, чем в устной. Для увеличения объема обучающих данных могут применяться различные методы аугментации, одним из широко используемых методов является частичный автоматический перевод, однако он применим, если существуют системы автоматического перевода для рассматриваемых языков. Так, пословный перевод с путунхуа на тайваньский диалект применялся в работах [4, 5]. В [6] аугментация включала случайную замену слов на перевод и применение метода на основе теории ограничения эквивалентности. В [7] применен схожий подход с грамматическими ограничениями, а в [8] генерация осуществлялась заменой слов на перевод либо объединением выбранных случайным образом предложений из исходного и параллельного (переведенного) корпуса с ограничением по близости векторных представлений слов. Также аугментация может выполняться путем искусственной генерации текста с помощью искусственных нейронных сетей [9, 10]. Кроме того, могут использоваться методы аугментации, применяемые для расширения объема моноязычных текстов, например, случайная замена, вставка, удаление слов или символов, контекстная аугментация и др. [11–13]. В работе [14], проведенной на материале арабских диалектных и англоязычных данных с переключением кодов, показано, что объединение систем с разной архитектурой позволяет повысить точность распознавания речи с переключением кода. В [15] приведено одно из первых экспериментальных исследований по моделированию переключения кодов между русским и казахским языками, показавшее, что одновременное билингвальное обучение может заметно повысить качество распознавания речи при наличии переключения кодов.

В работе [3] представлено сравнение моделей для распознавания речи с переключением кодов для различных языков. В зависимости от специфики языков и объемов обучающих данных полученные значения — показателя процента неправильно распознанных слов (Word Error Rate, WER) [16] — находятся в диапазоне от 22 % до 48,38 %.

При анализе научных публикаций по теме настоящей работы была обнаружена только одна работа [17], посвященная применению современных технологий для моделирования распознавания речи с переключением языков на материале малоресурсных языков России. В [17] исследован бесермянский диалект удмуртского языка, но при этом детальное рассмотрение методов аугментации представлено не было. Важно отметить, что для ливвиковского наречия карельского языка не имеется билингвальных текстовых баз данных с русским языком. При этом выполнение аугментации осложнено отсутствием систем автоматического перевода с карельского. В настоящей работе представлена методика аугментации текстовых данных для обучения языковых моделей, учитывающих переключение кода.

Обучение акустической модели карельско-русской речи

Акустическое моделирование карельской речи было выполнено путем дообучения предварительно обученной многоязычной интегральной модели Wav2Vec2-BERT 2.0[1]. Дообучение осуществлялось с применением фреймворка Transformers [18]. При дообучении модели использовались следующие параметры: регуляризация исключением весов механизма внимания — 0,94; регуляризация исключением латентных весов — 0,047; вероятность маски по времени — 0,082; максимальное количество шагов обучения — 10 000; логгирование каждые 500 шагов и оптимизатор AdamW со сниженной точностью 8 бит; коэффициент скорости обучения — 0,00002; размер батчей — 2 с 16 шагами накопления градиента и затуханием весов — 0,1. Материалом для дообучения модели послужили данные двух баз данных карельской речи: «База данных аннотаций речевых записей на карельском языке (AnKaS — Database of Annotations of Karelian Speech Recordings)» и «Речевая база данных с переключением кодов карельский-русский (KarRusCoS — Speech Database with Karelian-Russian Code-Switching)». Характеристики выбранных баз данных представлены в табл. 1, более подробное описание приведено в работах [19–21].

Речевые данные были разделены на обучающую, отладочную и тестовую выборки в соотношении примерно 8:1:1; речевые данные из базы данных AnKaS использовались только для обучения, а из KarRusCoS для обучения, отладки и тестирования. Данные для отладки и тестирования были выбраны случайным образом так, чтобы голоса одних и тех же дикторов не встречались в разных выборках. Была выполнена аугментация обучающей части речевой базы данных путем изменения темпа речи и частоты основного тона, а также одновременного изменения темпа речи и частоты основного тона. Кроме того, к обучающим данным были добавлены речевые данные на русском языке в объеме 6 ч, которые были взяты из речевой базы данных, описанной в [22], с расшифровками транслитерированными на латинице.

[1] [Электронный ресурс]. Режим доступа: https://huggingface.co/docs/transformers/model_doc/wav2vec2-bert (дата обращения: 20.07.2026).

Таблица 1. Характеристики баз данных карельской речи [19–21]
Table 1. Features of spoken Karelian corpora [19–21]

Параметр	Значение	
	AnKaS	KarRusCoS
Общее количество дикторов	17 (7 мужчин, 10 женщин)	40 (16 мужчин, 24 женщин)
Длительность аннотированных речевых данных, ч	4,5	3
Количество фраз	4385	3012
Количество словоупотреблений	32 037	22 355
Количество уникальных слов	9117	7091
Процент встраиваемого кода, %	1	28
Число слов с внутрисловным переключением кода, %	менее 1	6

Обучение языковой модели карельско-русской речи

Набор текстовых данных для обучения модели языка был сформирован из двух источников. Первый источник составляли расшифровки обучающей части речевых данных, в которых содержались 5855 фраз общим объемом в 47 тыс. словоупотреблений. Поскольку этот объем данных для обучения модели языка недостаточен, дополнительно была использована вторая база данных объемом в 5 млн словоупотреблений, содержащая тексты из книг, периодических изданий и открытых интернет-источников [23]. Эта текстовая база данных, в основном, является моноязычной и содержит мало примеров переключения на русский язык, поэтому была выполнена ее аугментация путем автоматического перевода части слов на русский язык. Для этого были разработаны правила перевода с карельского языка на русский, учитывающие морфонологические особенности переводимых слов. Дополнительно были разработаны правила для автоматической генерации слов с внутрисловным переключением кода.

Внутрисловное переключение кода представляет собой морфологическую адаптацию русских слов путем добавления карельских аффиксов к русской основе слова (например, для образования формы винительного падежа к русской основе «*учили*щ-» добавляется карельский аффикс «-*ап*», образуя словоформу «*училищап*»). Для разработки правил генерации слов с внутрисловным переключением кода были проанализированы расшифровки из базы данных KarRusCoS и определены аффиксы, которые наиболее часто используются при переключении кода, а также сформулированы правила морфологической адаптации карельских слов к русскому коду. Разработанные правила были применены к словарю русских слов, сформированному ранее для системы распознавания русской речи [24]. В результате сгенерирован список слов с внутрисловным переключением кода, который в дальнейшем использовался при создании модели языка. Подробнее процесс моделирования внутрисловного переключения кода описан в работе [21].

Процедура аугментации при помощи перевода осуществлялась по схеме, показанной на рис. 1. Поскольку в базе данных KarRusCoS количество русских слов составляет 28 %, в карельских текстах случайным образом было выбрано 28 % слов. Для этих слов с помощью словаря VepKar [25] были определены грамматические признаки, затем эти слова были приведены к словарной форме и переведены на русский язык (для перевода также использовался словарь VepKar). Если слово могло иметь разные наборы грамматических признаков или различные варианты перевода на русский язык, то выбор осуществлялся случайным образом, в результате отобран 21 % слов. Эти слова были преобразованы в слова с внутрисловным переключением кода в соответствии с их грамматическими признаками и разработанными правилами формирования таких слов. Для оставшихся переведенных лексем, не вошедших в 21 %, на основе их грамматических признаков, извлеченных из исходного контекста, генерировались соответствующе оформленные русские словоформы или словосочетания с тем же значением.

Аугментация посредством перевода выполнялась в несколько итераций (от 1 до 5) с разной инициализацией генератора случайных чисел. Начиная со второй итерации, к тексту добавлялись только новые предложения. Текстовый материал, взятый из расшифровок, не подвергался аугментации путем перевода, поскольку содержал переключения на русский язык. Для увеличения объема обучающих текстовых данных была выполнена аугментация по методике Easy Data Augmentation (EDA), которая включала в себя удаление, добавление, перемену мест и замену случайных слов. К тексту добавлялись только новые предложения. Данная аугментация была применена ко всей текстовой базе данных, включая текст, полученный за счет частичного перевода. Полученный таким образом текстовый материал далее использовался для обучения триграммной языковой модели, которое было выполнено с помощью программных средств SRILM [26]. Словарь языковых моделей был сформирован двумя способами: включением в словарь только слов, встретившихся в текстовых данных, при этом в словарь были включены все слова из расшифровок аудиоданных, карельские слова из текстового материала, встретившиеся в тексте более одного раза, и все русские (т. е. полученные путем перевода) слова (включая слова с внутрисловным переключением кода) из текстового материала (в дальнейшем такой словарь будет называться неполным); включением в словарь, помимо встретившихся в текстовой базе

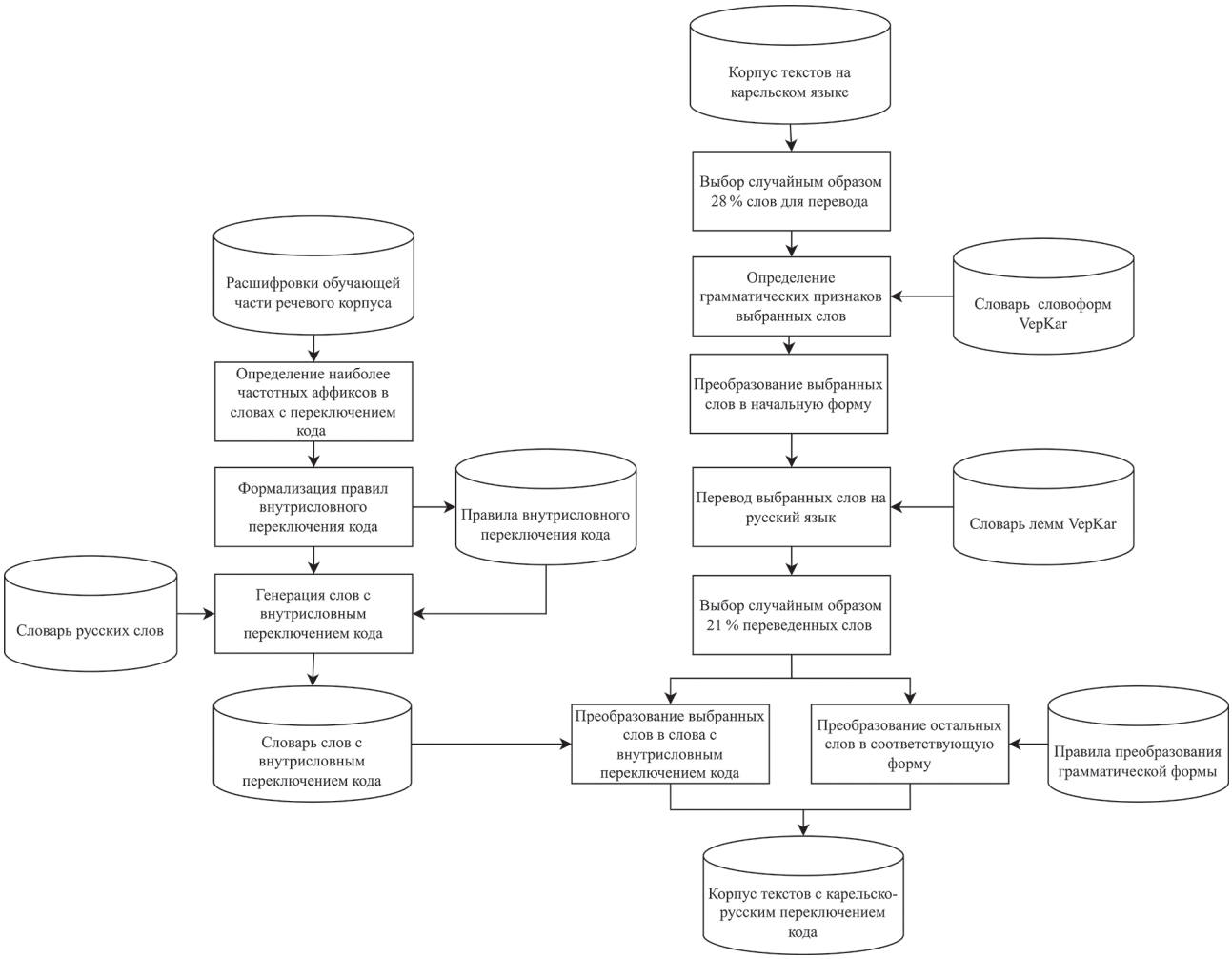

Рис. 1. Схема генерации текста с карельско-русским переключением кода
Fig. 1. Scheme of the generation of a Karelian-Russian code-switched text

данных, всех сгенерированных слов с внутрисловным переключением кода, а также словаря русских слов (объемом в 150 тыс. слов), созданного ранее для русскоязычной системы распознавания речи (в дальнейшем такой словарь будет называться полным).

Было обучено несколько статистических языковых моделей: на исходной текстовой базе данных (без аугментации) с неполным словарем; на аугментированных за счет частичного перевода текстовых данных с неполным словарем; на аугментированных за счет частичного перевода текстовых данных с полным словарем; на аугментированных как за счет частичного перевода, так и за счет EDA аугментации текстовых данных с полным словарем.

Также была выполнена линейная интерполяция языковых моделей карельского языка с созданной ранее моделью русского языка.

Архитектура системы распознавания речи с переключением кода с ливвиковского наречия карельского языка на русский

Разработанные акустические и языковые модели были применены для создания прототипа системы автоматического распознавания речи с переключением кода между карельским и русским языками. Структура системы представлена на рис. 2.

Система позволяет распознавать речь как с микрофона, так и из предварительно записанных файлов в формате WAV. Поддерживаемая частота дискретизации аудиозаписей — 16 000 Гц, разрешающая способность — 16 бит на цифровой отсчет. Вначале выполняется нормализация речевого сигнала, опционально производится идентификация активности голоса Voice Activity Detection, затем осуществляется декодирование речевого сигнала и выбор гипотезы распознавания с наибольшей оценкой вероятности, определенной с использованием акустической и языковой моделей. Система реализована в виде приложения с помощью библиотеки Gradio и может быть предоставлена по запросу.

Экспериментальные исследования разработанной системы распознавания карельской речи

В ходе экспериментальных исследований была выполнена оценка разработанной системы по показателю WER [16]. Эксперименты проводились в несколько

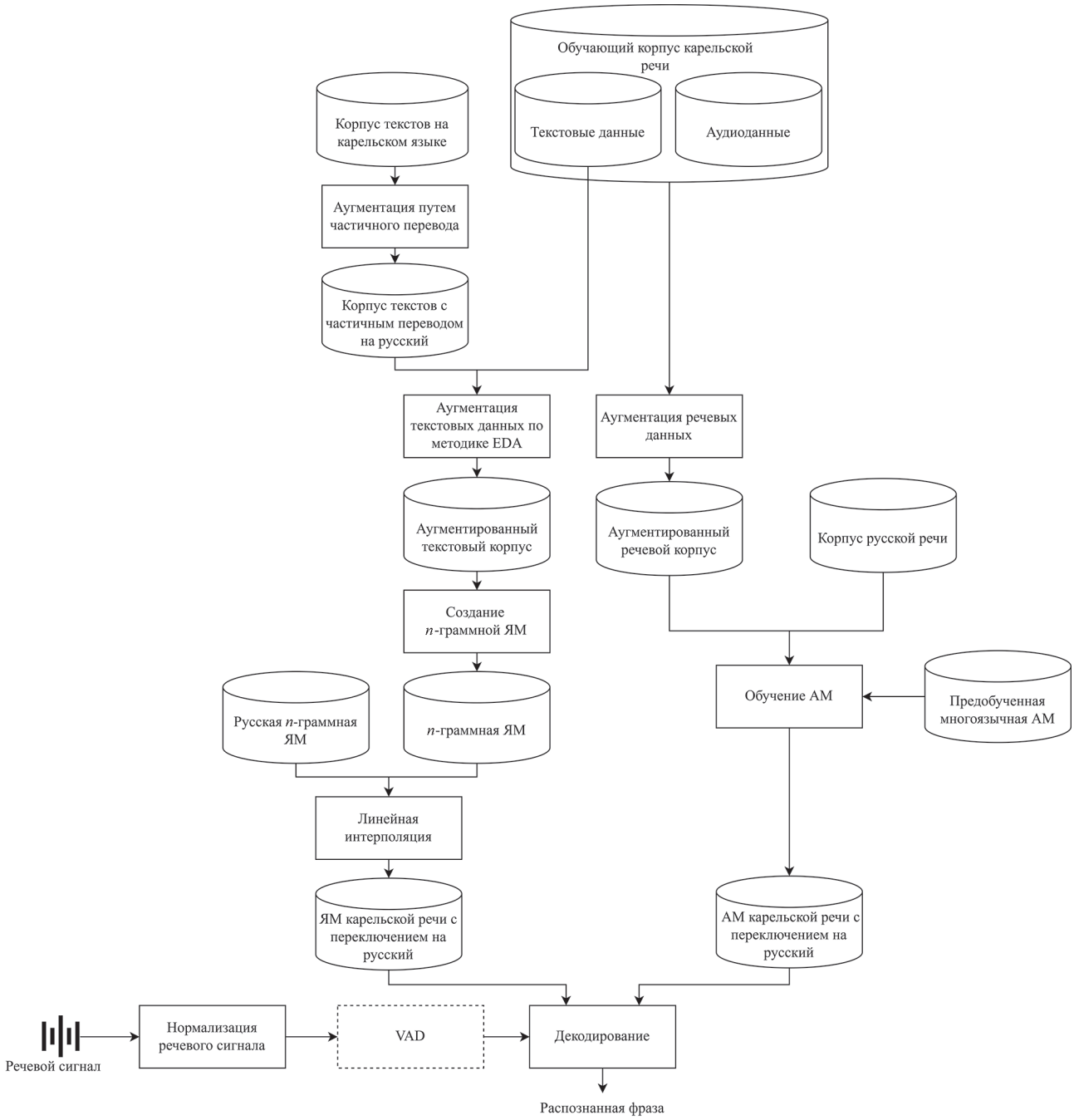

Рис. 2. Структура системы распознавания карельской речи с переключением на русский язык.
ЯМ — языковая модель; АМ — акустическая модель; VAD — Voice Activity Detection

Fig. 2. Structure of a recognition system for mixed Karelian-Russian speech.
ЯМ — language model; AM — acoustic model; VAD — Voice Activity Detection

этапов: вначале были проведены эксперименты с применением с базовой языковой модели, обученной на исходных текстах на карельском языке и расшифровках обучающей части речевой базы данных. При этом на отладочной части речевой базы данных значение WER составило 28,01 %, на тестовой — 30,83 %. Затем были проведены эксперименты с использованием языковых моделей, обученных на текстовых данных, полученных после применения аугментации за счет частичного перевода, выполненной с разным числом итераций, и с применением словарей разного объема. После этого была применена языковая модель, обученная на текстах, полученных с использованием методики EDA. Результаты проведенных экспериментов представлены в табл. 2.

Использование аугментации путем частичного перевода и полного словаря позволило снизить значение WER до 26,4 % на отладочном наборе данных, и до 29,54 % — на тестовом. Дополнительное небольшое снижение WER было достигнуто при выполнении не-

Таблица 2. Результаты распознавания речи на карельском языке с переключением на русский язык с применением различных моделей карельского языка

Table 2. Recognition results for Karelian speech with code-switching into Russian using different Karelian language models

Текстовый материал для обучения		Значение WER на корпусах, %			
		отладочном		тестовом	
		словарь содержит только слова из текстовой базы данных	полный словарь	словарь содержит только слова из текстовой базы данных	полный словарь
Расшифровки и переведенный текст	одна итерация	26,72	26,40	29,63	29,54
	две итерации	26,56	26,36	29,59	**29,52**
	три итерации	26,56	26,38	29,72	29,63
	четыре итерации	26,52	26,36	29,74	29,67
	5 итераций	26,51	26,34	29,74	29,67
	5 итераций + аугментация по методике EDA	—	**26,20**	—	29,70

Примечание: жирным выделены лучшие результаты.

скольких итераций аугментации. Наилучшие результаты на отладочном наборе данных были достигнуты на 5 итерации, однако увеличение числа итераций привело лишь к незначительному снижению WER, поэтому больше 5 итераций аугментации не проводилось. Отметим, что начиная с третьей итерации при экспериментах на тестовом наборе данных значение WER начало расти.

В табл. 2 также представлены результаты по применению языковой модели, обученной на тексте с аугментацией по методике EDA. Данная аугментация осуществлялась на текстовых данных, полученных на 5 итерации, поскольку именно такое число итераций дало наилучшие результаты на отладочной базе данных. Как видно из табл. 2, данная аугментация также позволила улучшить результаты распознавания. Таким образом, наилучшие результаты на отладочном наборе данных были получены при использовании языковой модели, обученной на текстовом материале, сформированном за счет аугментации путем перевода отдельных слов на русский язык, выполненной 5 раз, и аугментации по методике EDA. При этом значение WER на отладочном наборе данных составило 26,2 %, на тестовом — 29,7 %.

В заключение, были проведены эксперименты с использованием языковой модели, полученной путем линейной интерполяции с русской моделью. Наилучшие результаты были достигнуты при применении коэффициента интерполяции, равного 0,7, при этом на отладочном наборе данных значение WER составило 25,82 %, на тестовом — 29 %. Таким образом, относительное снижение WER по сравнению с использованием базовой языковой модели на отладочном наборе данных составило 7,82 %, на тестовом — 5,94 %, что показывает эффективность разработанной методики аугментации текстовых данных для обучения модели карельского языка с переключением кода на русский язык.

Полученные результаты находятся на уровне мировых результатов для других малоресурсных языков с переключением кода. Различия в полученных результатах на отладочном и тестовом наборах данных могут быть связаны с тем, что объем используемых речевых данных был небольшим, при этом дикторы в обучающих, отладочных и тестовых данных не пересекались. Различные дикторы могут переключаться между языками по-разному, что объясняет тот факт, что для разных дикторов изменение показателя WER с увеличением числа итераций было различным.

Заключение

Представлена методика, при помощи которой предложено решение задачи создания системы автоматического распознавания речи для малоресурсного языка с переключением кодов. Анализ эффективности различных методов аугментации текстовых данных позволяет сделать следующие наблюдения. Частичный автоматический перевод карельских текстов на русский язык с последующей морфологической адаптацией заимствований обеспечивает снижение Word Error Rate (WER) относительно модели, обученной только на расшифровках. Дополнительное небольшое снижение WER достигается при выполнении нескольких итераций аугментации. Однако простое увеличение числа итераций аугментации в проведенных экспериментах не привело к пропорциональному улучшению результатов. Это может объясняться тем, что многократно аугментированные тексты начинают статистически отклоняться от реального распределения переключения кода.

Включение в словарь сгенерированных словоформ с внутрисловным переключением кода дает стабильное, но незначительное улучшение результатов. Данный эффект достаточно легко объясним и даже предсказуем, поскольку внутрисловное переключение составляет 6 % от общего числа словоупотреблений в базе данных KarRusCoS. Дальнейший прирост качества распознавания продемонстрировала линейная интерполяция карельской языковой модели с русской. Отметим, что разработанные правила морфонологической адаптации охватывают только наиболее частотные стратегии вну-

трисловного переключения и могут не отражать всего разнообразия языковых структур, встречающихся в реальной речи носителей карельского языка.

Предложенный подход, подразумевающий дообучение многоязычной модели в сочетании с текстовой аугментацией и интерполяцией языковых моделей, может быть успешно применен к иным малоресурсным языкам России. Тем не менее, следует иметь в виду, что для этого потребуется либо разработка аналогичных правил морфонологической адаптации для каждого целевого языка, либо статистическое описание конкретного языкового материала, не требующее описания на основе правил.

Литература

1. Mabokela K.R., Manamela M.J., Manaileng M. Modeling code-switching speech on under-resourced languages for language identification // Proc. of the 4th Workshop on Spoken Language Technologies for Under Resourced Languages (SLTU). 2014. P. 225–230.
2. Bhuvanagirir K., Kopparapu S.K. Mixed language speech recognition without explicit identification of language // American Journal of Signal Processing. 2012. V. 2. N 5. P. 92–97. doi: 10.5923/j.ajsp.20120205.02
3. Agro M.T., Kulkarni A.A., Kadaoui K., Talat Z., Aldarmaki H. Code-switching in End-to-End automatic speech recognition: A systematic literature review // arXiv. 2025. arXiv:2507.07741. doi: 10.48550/arXiv.2507.07741
4. Hsieh I.-T., Wu C.-H., Wang C.-H. Acoustic and textual data augmentation for code-switching speech recognition in under-resourced language // Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2020. P. 302–307.
5. Hamed I., Habash N., Abdennadher S., Vu N.T. Investigating lexical replacements for Arabic-English code-switched data augmentation // Proc. of the 6th Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT). 2023. P. 86–100. doi: 10.18653/v1/2023.loresmt-1.7
6. Hussein A., Chowdhury S.A., Abdelali A., Dehak N., Ali A., Khudanpur S. Textual data augmentation for Arabic-English code-switching speech recognition // Proc. of the IEEE Spoken Language Technology Workshop (SLT). 2023. P. 777–784. doi: 10.1109/SLT54892.2023.10023313
7. Pratapa A., Bhat G., Choudhury M., Sitaram S., Dandapat S., Bali K. Language modeling for code-mixing: The role of linguistic theory based synthetic data // Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. V. 1. P. 1543–1553. doi: 10.18653/v1/P18-1143
8. Chuang S.-P., Sung T.-W., Lee H.-Y. Training code-switching language model with monolingual data // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 7949–7953. doi: 10.1109/ICASSP40776.2020.9053775
9. Yilmaz E., van Den Heuvel H., van Leeuwen D.A. Acoustic and textual data augmentation for improved ASR of code-switching speech // Proc. of the 19th Annual Conference of the International Speech Communication Association Interspeech. 2018. P. 1933–1937. doi: 10.21437/Interspeech.2018-52
10. Chang C.-T., Chuang S.-P., Lee H.-Y. Code-switching sentence generation by generative adversarial networks and its application to data augmentation // Proc. of the 20th Annual Conference of the International Speech Communication Association Interspeech. 2019. P. 554–558. doi: 10.21437/Interspeech.2019-3214
11. Şahin G.G. To augment or not to augment? A comparative study on text augmentation techniques for low-resource NLP // Computational Linguistics. 2022. V. 48. N 1. P. 5–42. doi: 10.1162/coli_a_00425
12. Wan Z., Wan X., Peng W., Li R. New datasets and controllable iterative data augmentation method for code-switching ASR error correction // Findings of the Association for Computational Linguistics: EMNLP. 2023. P. 8075–8087. doi: 10.18653/v1/2023.findings-emnlp.543
13. Kobayashi S. Contextual augmentation: Data augmentation by words with paradigmatic relations // Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018. V. 2. P. 452–457. doi: 10.18653/v1/N18-2072
14. Hamed I., Denisov P., Li C.-Y., Elmahdy M., Abdennadher S, Vu N.T. Investigations on speech recognition systems for low-resource

References

1. Mabokela K.R., Manamela M.J., Manaileng M. Modeling code-switching speech on under-resourced languages for language identification. *Proc. of the 4th Workshop on Spoken Language Technologies for Under Resourced Languages (SLTU)*, 2014, pp. 225–230.
2. Bhuvanagirir K., Kopparapu S.K. Mixed language speech recognition without explicit identification of language. *American Journal of Signal Processing*, 2012, vol. 2, no. 5, pp. 92–97. doi: 10.5923/j.ajsp.20120205.02
3. Agro M.T., Kulkarni A.A., Kadaoui K., Talat Z., Aldarmaki H. Code-switching in End-to-End automatic speech recognition: A systematic literature review. *arXiv*, 2025. arXiv:2507.07741. doi: 10.48550/arXiv.2507.07741
4. Hsieh I.-T., Wu C.-H., Wang C.-H. Acoustic and textual data augmentation for code-switching speech recognition in under-resourced language. *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 302–307.
5. Hamed I., Habash N., Abdennadher S., Vu N.T. Investigating lexical replacements for Arabic-English code-switched data augmentation. *Proc. of the 6th Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT)*, 2023, pp. 86–100. doi: 10.18653/v1/2023.loresmt-1.7
6. Hussein A., Chowdhury S.A., Abdelali A., Dehak N., Ali A., Khudanpur S. Textual data augmentation for Arabic-English code-switching speech recognition. *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 777–784. doi: 10.1109/SLT54892.2023.10023313
7. Pratapa A., Bhat G., Choudhury M., Sitaram S., Dandapat S., Bali K. Language modeling for code-mixing: The role of linguistic theory based synthetic data. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, vol. 1, pp. 1543–1553. doi: 10.18653/v1/P18-1143
8. Chuang S.-P., Sung T.-W., Lee H.-Y. Training code-switching language model with monolingual data. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7949–7953. doi: 10.1109/ICASSP40776.2020.9053775
9. Yilmaz E., van Den Heuvel H., van Leeuwen D.A. Acoustic and textual data augmentation for improved ASR of code-switching speech. *Proc. of the 19th Annual Conference of the International Speech Communication Association Interspeech*, 2018, pp. 1933–1937. doi: 10.21437/Interspeech.2018-52
10. Chang C.-T., Chuang S.-P., Lee H.-Y. Code-switching sentence generation by generative adversarial networks and its application to data augmentation. *Proc. of the 20th Annual Conference of the International Speech Communication Association Interspeech*, 2019, pp. 554–558. doi: 10.21437/Interspeech.2019-3214
11. Şahin G.G. To augment or not to augment? A comparative study on text augmentation techniques for low-resource NLP. *Computational Linguistics*, 2022, vol. 48, no. 1, pp. 5–42. doi: 10.1162/coli_a_00425
12. Wan Z., Wan X., Peng W., Li R. New datasets and controllable iterative data augmentation method for code-switching ASR error correction. *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 8075–8087. doi: 10.18653/v1/2023.findings-emnlp.543
13. Kobayashi S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, vol. 2, pp. 452–457. doi: 10.18653/v1/N18-2072
14. Hamed I., Denisov P., Li C.-Y., Elmahdy M., Abdennadher S, Vu N.T. Investigations on speech recognition systems for low-resource

dialectal Arabic–English code-switching speech // Computer Speech & Language. 2022. V. 72. P. 101278. doi: 10.1016/j.csl.2021.101278
15. Ubskii D., Matveev Y., Minker W. Impact of using a bilingual model on Kazakh-Russian code-switching speech recognition // CEUR Workshop Proceedings. 2020. V. 2590. P. 1–6.
16. Карпов А.А., Кипяткова И.С. Методология оценивания работы систем автоматического распознавания речи // Известия высших учебных заведений. Приборостроение. 2012. Т. 55. № 11. С. 38–43.
17. Arkhangelskiy T. Low-resource ASR with an augmented language model // Proc. of the 7th International Workshop on Computational Linguistics of Uralic Languages. 2021. P. 40–46.
18. Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., et al. Transformers: State-of-the-art natural language processing // Proc. of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. P. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6
19. Кипяткова И.С., Кагиров И.А., Долгушин М.Д. Применение предварительно обученных многоязычных моделей для распознавания карельской речи // Информатика и автоматизация. 2025. Т. 24. № 2. С. 604–630. doi: 10.15622/ia.24.2.9
20. Кагиров И.А., Киселева К.О., Кипяткова И.С. Анализ переключения кодов в речи носителей карельского языка // Terra Linguistica. 2025. Т. 16. № 2. С. 22–40. doi: 10.18721/JHSS.16202
21. Kipyatkova I., Kiseleva K., Dolgushin M., Kagirov I. Modeling intra-word code-switching for Karelian ASR // Lecture Notes in Computer Science. 2025. V. 16188. P. 104–117. doi: 10.1007/978-3-032-07959-6_8
22. Kipyatkova I. Experimenting with hybrid TDNN/HMM acoustic models for Russian speech recognition // Lecture Notes in Computer Science. 2017. V. 10458. P. 362–369. doi: 10.1007/978-3-319-66429-3_35
23. Кипяткова И.С., Родионова А.П., Кагиров И.А., Крижановский А.А. Подготовка речевых и текстовых данных для создания системы автоматического распознавания карельской речи // Ученые записки Петрозаводского государственного университета. 2023. Т. 45. № 5. С. 89–98. doi: 10.15393/uchz.art.2023.924
24. Kipyatkova I., Karpov A. Lexicon size and language model order optimization for Russian LVCSR // Lecture Notes in Computer Science. 2013. V. 8113. P. 219–226. doi: 10.1007/978-3-319-01931-4_29
25. Boyko T., Zaitseva N., Krizhanovskaya N., Krizhanovsky A., Novak I., Pellinen N., et al. The open corpus of the Veps and Karelian languages: Overview and applications // KnE Social Sciences. 2022. P. 29–40. doi: 10.18502/kss.v7i3.10419
26. Stolcke A., Zheng J., Wang W., Abrash V. SRILM at sixteen: update and outlook // Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop. 2011. P. 5–9.

dialectal Arabic–English code-switching speech. *Computer Speech & Language*, 2022, vol. 72, pp. 101278. doi: 10.1016/j.csl.2021.101278
15. Ubskii D., Matveev Y., Minker W. Impact of using a bilingual model on Kazakh-Russian code-switching speech recognition. *CEUR Workshop Proceedings*, 2020, vol. 2590, pp. 1–6.
16. Karpov A.A., Kipyatkova I.S. Methodology for estimation of automatic speech recognition system performance. *Journal of Instrument Engineering*, 2012, vol. 55, no. 11, pp. 38–43. (in Russian)
17. Arkhangelskiy T. Low-resource ASR with an augmented language model. *Proc. of the 7th International Workshop on Computational Linguistics of Uralic Languages*, 2021, pp. 40–46.
18. Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., et al. Transformers: State-of-the-art natural language processing. *Proc. of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6
19. Kipyatkova I., Kagirov I., Dolgushin M. Use of pre-trained multilingual models for Karelian speech recognition. *Informatics and Automation*, 2025, vol. 24, no. 2, pp. 604–630. (in Russian). doi: 10.15622/ia.24.2.9
20. Kagirov I.A., Kiseleva K.O., Kipyatkova I.S. Code-switching analysis in Karelian language speakers' speech. *Terra Linguistica*, 2025, vol. 16, no. 2, pp. 22–40. (in Russian). doi: 10.18721/JHSS.16202
21. Kipyatkova I., Kiseleva K., Dolgushin M., Kagirov I. Modeling intra-word code-switching for Karelian ASR. *Lecture Notes in Computer Science*, 2025, vol. 16188, pp. 104–117. doi: 10.1007/978-3-032-07959-6_8
22. Kipyatkova I. Experimenting with hybrid TDNN/HMM acoustic models for Russian speech recognition. *Lecture Notes in Computer Science*, 2017, vol. 10458, pp. 362–369. doi: 10.1007/978-3-319-66429-3_35
23. Kipyatkova I.S., Rodionova A.P., Kagirov I.A., Krizhanovsky I.A. Speech and text data preparation for development of an automatic speech recognition system for the Karelian language. *Proceedings of Petrozavodsk State University*, 2023, vol. 45, no. 5, pp. 89–98. (in Russian). doi: 10.15393/uchz.art.2023.924
24. Kipyatkova I., Karpov A. Lexicon size and language model order optimization for Russian LVCSR. *Lecture Notes in Computer Science*, 2013, vol. 8113, pp. 219–226. doi: 10.1007/978-3-319-01931-4_29
25. Boyko T., Zaitseva N., Krizhanovskaya N., Krizhanovsky A., Novak I., Pellinen N., et al. The open corpus of the Veps and Karelian languages: Overview and applications. *KnE Social Sciences*, 2022, pp. 29–40. doi: 10.18502/kss.v7i3.10419
26. Stolcke A., Zheng J., Wang W., Abrash V. SRILM at sixteen: update and outlook. *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 5–9.

Авторы

Кипяткова Ирина Сергеевна — кандидат технических наук, доцент, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 36018436400, https://orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su

Долгушин Михаил Дмитриевич — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 58743617900, https://orcid.org/0000-0002-4344-2330, dolgushin.m@iias.spb.su

Киселева Ксения Олеговна — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 60000649000, https://orcid.org/0009-0001-5572-3842, kiseleva.k@iias.spb.su

Кагиров Ильдар Амирович — научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 25121369400, https://orcid.org/0000-0003-1196-1117, kagirov@iias.spb.su

Authors

Irina S. Kipyatkova — PhD, Associate Professor, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 36018436400, https://orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su

Mikhail D. Dolgushin — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 58743617900, https://orcid.org/0000-0002-4344-2330, dolgushin.m@iias.spb.su

Kseniia O. Kiseleva — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 60000649000, https://orcid.org/0009-0001-5572-3842, kiseleva.k@iias.spb.su

Ildar A. Kagirov — Scientific Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 25121369400, https://orcid.org/0000-0003-1196-1117, kagirov@iias.spb.su

Статья поступила в редакцию 10.04.2026
Одобрена после рецензирования 24.06.2026
Принята к печати 22.07.2026

Received 10.04.2026
Approved after reviewing 24.06.2026
Accepted 22.07.2026