УДК 004.934
doi:10.31799/1684-8853-2023-3-16-25
EDN: YOLUUY

Научные статьи
Articles

Система автоматического распознавания карельской речи

И. С. Кипяткова[a], *канд. техн. наук, доцент, http://orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su*
И. А. Кагиров[a], *научный сотрудник, orcid.org/0000-0003-1196-1117*
[a]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *в последнее время растет число исследований, посвященных автоматической обработке малоресурсных языков. Отсутствие или малый объем обучающих данных является существенным препятствием в развитии речевых технологий для подобных языков.* ***Цель:*** *разработать систему автоматического распознавания речи на карельском языке.* ***Результаты:*** *представлена система автоматического распознавания карельской речи. Обучены акустические модели на основе искусственных нейронных сетей с временными задержками и скрытых марковских моделей. Обучение осуществлялось на речевом корпусе, составленном из записей радиопередач и аудиоданных, полученных путем аугментации. Модель карельского языка обучалась как на письменных текстах, так и на расшифровках обучающей части речевого корпуса. Во время обучения исследовались различные коэффициенты для интерполяции языковой модели, обученной на расшифровках, с моделью языка, обученной на письменных текстах. В ходе экспериментов по распознаванию карельской речи лучший результат по показателю количество неправильно распознанных слов составил 25,81 %, что сопоставимо с общим уровнем распознавания речи для других малоресурсных языков. Собран обучающий набор данных, который включает звукозаписи на карельском языке с расшифровками, а также текстовый корпус.* ***Практическая значимость:*** *полученные решения могут играть роль в создании автоматических систем распознавания не только карельского, но и других малоресурсных языков. Разработанная система поможет исследователям карельского языка, предоставляя эффективный инструмент для записи и обработки карельского языкового материала.*

Ключевые слова *— малоресурсные языки, автоматическое распознавание речи, карельский язык, искусственные нейронные сети с временной задержкой.*

Для цитирования: Кипяткова И. С., Кагиров И. А. Система автоматического распознавания карельской речи. *Информационно-управляющие системы*, 2023, № 3, с. 16–25. doi:10.31799/1684-8853-2023-3-16-25, EDN: YOLUUY
For citation: Kipyatkova I. S., Kagirov I. A. Automatic speech recognition system for Karelian. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2023, no. 3, pp. 16–25 (In Russian). doi:10.31799/1684-8853-2023-3-16-25, EDN: YOLUUY

Введение

Современные технологии обработки естественного языка активно развиваются и находят применение в различных областях, включая автоматическое распознавание речи. Однако отсутствие развитых речевых технологий для малоресурсных языков остается актуальной проблемой [1]. В связи с этим авторы настоящего исследования предлагают ее решение для карельского языка.

Карельский язык — это язык из финно-угорской языковой семьи, который в настоящее время используется на территории Республики Карелия (Россия). Он относится к прибалтийско-финской группе и близок к вепсскому и ижорскому языкам [2] как генетически, так и типологически. Карельский язык также имеет типологическое сходство с другими региональными языками, в том числе с финским и эстонским. Карельский язык относится к уязвимым языкам, в частности из-за уменьшения числа носителей.

Создание системы автоматической обработки естественного языка составляет сложную и нетривиальную задачу, и одним из важнейших условий ее реализации является наличие обучающих наборов данных. Так как объем электронных языковых ресурсов для карельского языка невелик, то этот вопрос является принципиальным. Также важным моментом при разработке системы автоматического распознавания речи является выбор оптимального алгоритма обучения системы. В рамках данной статьи рассмотрен ряд систем распознавания речи, созданных как для малоресурсных, так и для генетически близких (прибалтийско-финских) языков, и проводится сравнение нескольких подходов к автоматическому распознаванию.

Настоящая работа может иметь определенное значение с точки зрения разработки информационного обеспечения для малоресурсных языков [3–5], поскольку полученные в ходе работы решения помогут при создании автоматических систем распознавания речи не только для карельского, но и других языков.

Аналитический обзор систем автоматического распознавания речи для прибалтийско-финских и малоресурсных языков

Несколько работ, появившихся за последнее время, позволяют оценить эффективность подходов к распознаванию речи, применявшихся для родственных карельскому языков, — финского и эстонского. Несмотря на то, что оба этих языка не являются малоресурсными и примененные методики не могут быть напрямую перенесены на карельский материал, информация о методах является достаточно ценной в силу схожести языков.

В статье [6] описан речевой корпус, состоящий из записей заседаний Парламента Финляндии и снабженный автоматическими расшифровками. Авторы провели анализ качества распознавания речи на этом корпусе, используя несколько метрик и методов оценки. Сравнивались различные модели распознавания речи, включая скрытые марковские модели (СММ), гибридные модели, объединяющие искусственные нейронные сети (ИНС) и СММ — ИНС/СММ, а также модели с архитектурой кодер-декодер с механизмом внимания (Encoder-Decoder with the Attention Mechanism). Из проведенных экспериментов следует, что наилучшие результаты достигаются при использовании гибридных ИНС/СММ моделей. Кроме того, было выяснено, что наличие предобученных на других языках моделей может привести к улучшению качества распознавания речи на финском языке.

В работе [7] представлена система распознавания речи на эстонском языке, обученная на речевом корпусе объемом 268,5 ч. Акустическое моделирование осуществлялось факторизованными ИНС с временными задержками (Factorized Time Delay Neural Networks, TDNN-F) совместно с СММ. Для обучения акустической модели применялся критерий безрешеточной максимизации взаимной информации (Lattice-Free Maximum Mutual Information) [8]. Гиперпараметры настройки обучения взяты из библиотеки Kaldi Switchboard. Декодирование речи выполнялось с использованием 4-граммной модели языка, кроме того, осуществлялась переоценка гипотез распознавания при помощи нейросетевой модели языка. В результате экспериментов количество неправильно распознанных слов (Word Error Rate, WER) составило 8,1 % на тестовом наборе данных, содержащем записи ток-шоу и телефонных интервью, 12,9 % — при распознавании записей с конференций и 22,7 % – на зашумленных записях.

Развитие описанной выше системы распознавания эстонской речи представлено в работе [9]. Авторы расширили корпус эстонской речи до 761 ч и исследовали интегральный (end-to-end) подход с использованием предобученной модели wav2vec2.0 [10]. Авторам удалось снизить значение WER до 6,9 % на тестовом наборе данных, собранных из записей радио- и телепередач. Следует отметить, что при большом объеме обучающих данных интегральные системы показывают лучшую производительность с точки зрения скорости и точности распознавания речи, однако для их обучения требуется существенно больший объем данных, и при недостатке обучающих данных точность таких систем ниже, чем систем, построенных из отдельных компонентов [11].

В статье [12] отмечается, что в целом интегральный подход к автоматическому распознаванию речи [13, 14] оказывается рациональным для бесписьменных или находящихся под угрозой исчезновения языков, поскольку зачастую сбор переводов на высокоресурсный язык оказывается легче, чем транскрибирование записей исходного языка [15]. Тем не менее создание высококачественной интегральной системы с небольшим количеством исходных параллельных данных представляет собой проблему при отсутствии доступа к параллельным корпусам языковых данных. В том случае, если дополнительных ресурсов для исходного языка нет, хорошим подходом оказывается применение метода переноса знаний (transfer learning), суть которого сводится к использованию результатов обучения родительской модели языка, полученных на большом наборе данных, для инициализации весов в дочерней модели, обученной на данных целевого малоресурсного языка. Например, в работе [16] продемонстрировано, что предварительное обучение системы автоматического распознавания речи на материале английского и французского языков позволяет существенно улучшить точность распознавания для испанского языка. Применение подобного подхода для малоресурсных языков описано в работах [17, 18].

Несмотря на широкое распространение интегрального подхода к распознаванию речи, при разработке систем распознавания для малоресурсных языков чаще всего используют стандартный подход, при котором система строится из трех отдельных компонентов (моделей): акустической, языковой и лексической (словаря), — поскольку при этом подходе требуется меньше данных для обучения модели. Такой подход применялся, например, в исследовании [19] для распознавания речи на малоресурсном сингальском языке (о. Шри-Ланка). Полученные авторами результаты показывают, что применение гибридных ИНС/СММ акустических моделей превосходит использование статистических СММ на

7,48 % по показателю WER на тестовом наборе данных. Наименьшее значение WER (35,16 %) получено при использовании архитектуры ИНС с временными задержками (Time Delay Neural Network, TDNN) [20] для создания акустической модели.

Авторами работы [21] представлены результаты экспериментов по многоязычному распознаванию речи на малоресурсных языках (10 языков из набора, предложенного в рамках соревнования OpenASR20 (https://sat.nist.gov/openasr20), а также на североамериканских языках кри и инуитских. В работе исследовалось применение TDNN-F в гибридных ИНС/СММ акустических моделях и показано, что в этом случае значения WER ниже, чем при применении двунаправленных ИНС с долгой кратковременной памятью (LSTM). Тем не менее представленные значения WER достаточно большие и варьируются от 48 до 69,6 % в зависимости от языка [21]. Похожий результат был получен для сомалийского языка [22].

В различных исследованиях, проводимых для русского языка, также было установлено, что использование гибридных акустических моделей на основе TDNN по точности распознавания речи превосходит использование СММ с гауссовыми смесями, а также гибридных ИНС/СММ с другими архитектурами нейронных сетей [23, 24].

Практически обязательным этапом создания системы распознавания речи для малоресурсных языков является аугментация данных — метод создания дополнительных данных путем изменения (модификации) собранных обучающих данных. К распространенным методам аугментации речевых данных относятся изменение частоты основного тона, темпа речи, преобразование голоса, изменение спектрограммы, синтез речи [25–27]. Для расширения набора текстовых данных может выполняться контекстная аугментация, аугментация на основе замены символов или слов, а также обратный перевод [28–30]. Чаще всего наилучшие результаты дает применение сразу нескольких методов аугментации [31]. Подробный обзор методов, применяемых при разработке систем автоматического распознавания речи для решения проблемы недостаточного количества обучающих данных, представлен в работе [32].

По результатам проведенного в ходе данного исследования обзора было принято решение использовать стандартный подход к построению системы распознавания речи для карельского языка, гибридные ИНС/СММ акустические модели, а также методы аугментации речевых данных для расширения обучающего речевого корпуса.

Речевые и текстовые данные для обучения системы распознавания карельской речи

Обучение системы автоматического распознавания речи производится с использованием двух корпусов: речевого и текстового. В качестве речевого корпуса в рамках выполненного исследования использовались записи радиопередач на карельском языке, представляющие собой интервью с двумя и более дикторами (в общей сложности 15, из них 6 мужчин и 9 женщин). Были проведены обработка и аннотирование данного корпуса. Одна из проблем, возникших в ходе работы, связана с наложениями речи, т. е. с фразами, содержащими одновременную речь двух дикторов. Другую проблему составили фоновые шумы, которые сильно ухудшили качество записей. Хотя для создания корпуса использовались только записи студийного качества, фоновый шум в некоторых случаях все же имел место. Все записи с фоновым шумом и наложениями речи были удалены из базы данных.

Одной из особенностей современного карельского языка является переключение кодов [33], которое представляет собой спонтанный переход говорящего с одного языка или диалекта на другой. В данное время среди носителей карельского языка в России распространено карельско-русское двуязычие [34], поэтому переход на русский язык и обратно вполне естественен. Поскольку в настоящем исследовании не ставилась задача обработки явления переключения кодов, фразы, содержащие слова на русском языке, не были включены в корпус.

Итоговый объем речевого корпуса составил 3,5 ч, общее число записанных фраз — 3819. Корпус разбит на обучающую и тестовую части в соотношении 9:1 (табл. 1).

Дополнительно для расширения обучающей части речевого корпуса проведена аугментация

■ ***Таблица 1.*** Характеристики речевого корпуса
■ ***Table 1.*** Speech corpus features

Параметр	Значение
Количество дикторов	15 (6 муж., 9 жен.)
Общая продолжительность слитной речи	3,5 ч
Общий объем данных	2,2 Гб
Количество фраз	3819
Частота дискретизации аудио	16 000 Гц
Квантование сигнала	16 бит
Соотношение обучающей/тестовой части	9:1

■ ***Таблица 2.*** Характеристики текстового корпуса
■ ***Table 2.*** Text corpus features

Параметр	Значение
Общий объем	5 млн словоупотреблений
Соотношение объема материала по источникам, %:	
книги	22,6
периодические издания	73,1
тексты из корпуса «ВепКар»	3,8
расшифровки аудиоматериала	0,5
Нормализация	Сегментация текста по предложениям Замена заглавных букв на строчные Удаление знаков препинания

данных, для чего был использован инструментарий Sox (http://sox.sourceforge.net/sox.html), с помощью которого выполнено изменение темпа речи и высоты голоса диктора, что позволило увеличить объем обучающего речевого материала в два раза.

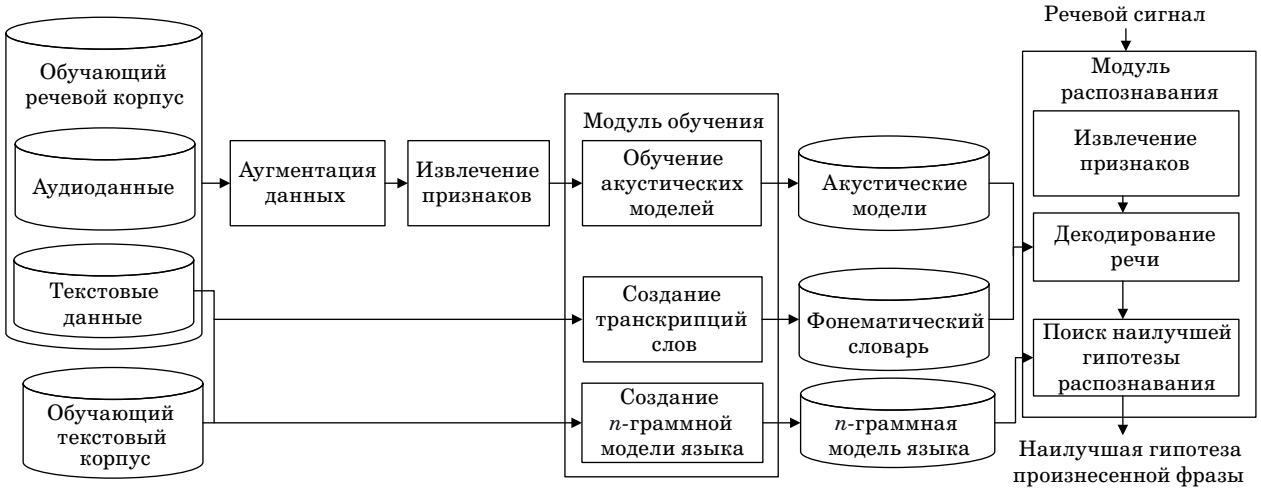
Собранный в ходе данного исследования текстовый корпус включает материалы из таких источников, как печатные издания, периодика на ливвиковском наречии, тексты из открытого корпуса на вепсском и карельском языках («ВепКар» – http://dictorpus.krc.karelia.ru/ru), а также расшифровки аудиозаписей обучающей части речевого корпуса (табл. 2). Все тексты приведены в формат .txt. Текстовая часть базы данных получена с частичным применением полуавтоматического распознавания текста.

Более подробно процесс сбора и обработки речевого и текстового корпусов описан в работе [35].

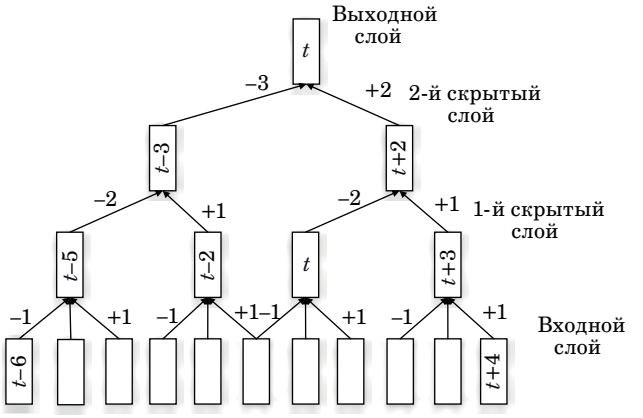
Система распознавания карельской речи

Структура системы распознавания карельской речи представлена на рис. 1.

В качестве акустической модели использовалась гибридная модель с TDNN, аналогичная архитектуре, показавшей наилучшие результаты для русской речи [24]. Обучение осуществлялось с помощью библиотеки nnet2 для рецепта swbd (s5c) инструментария Kaldi [8], при этом был применен стандартный метод обратного распространения ошибки (backpropagation), в качестве функции потерь применялась перекрестная энтропия. Для сокращения времени обучения использовалась технология увеличения скорости обучения, которая предполагает, что веса элементов в скрытом слое обучаются только на некоторых временных шагах, а не на каждом временном шаге [20]. На рис. 2 показан пример архитектуры для TDNN при временном контексте [–6, 4], интервал состоит из целых чисел,



■ ***Рис. 1.*** Структура системы распознавания карельской речи
■ ***Fig. 1.*** Outline of the Karelian speech recognition system



■ ***Рис. 2.*** Пример архитектуры TDNN с применением технологии объединения для временного контекста

■ ***Fig. 2.*** An example of TDNN architecture with sub-sampling

соответствующих временным шагам. Входной слой объединяет фреймы в интервале {–1, 0, 1} (более компактно это можно записать как [–1, 1]). Для скрытого слоя объединение осуществляется для временных шагов {–2, 1}, это означает, что объединяются фрейм, располагающийся на временном контексте за два фрейма до текущего, и фрейм, располагающийся через один фрейм после текущего. На втором скрытом слое объединяется фрейм, располагающийся за три фрейма до текущего, и фрейм, располагающийся через два фрейма после текущего ({–3, 2}).

Входными данными для нейронной сети были мел-частотные кепстральные коэффициенты, при этом для адаптации к речи диктора к ним был добавлен 100-мерный *i*-вектор, использование которого, как показано в работе [36], позволяет снизить WER. В данной работе были обучены ИНС с применением активационной функции p-norm [37], которая вычисляется следующим образом:

$$y = \|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}},$$

где вектор **x** представляет небольшую группу входных данных (мини-батч). Величина p выбирается опытным путем, в работе [37] показано, что p=2 дает лучшие результаты. Выходными данными нейронной сети являются апостериорные вероятности контекстно-зависимых моделей фонем.

Для ИНС с активационной функцией p-norm вместо параметра размерности скрытого слоя используются два параметра: число входов и число выходов. Отношение числа входов к числу выходов должно быть целым числом. Обычно используется отношение 5 или 10 [37]. Были созданы ИНС с различным числом скрытых слоев, различным временным контекстом и различными индексами объединяемых элементов. В процессе обучения коэффициент скорости обучения уменьшался от 0,02 до 0,004 в течение 15 эпох, затем пять эпох обучение происходило с постоянным коэффициентом скорости обучения, равным 0,004.

В словарь системы распознавания вошли все слова из расшифровок обучающей части речевого корпуса и слова из остальных текстов, которые встретились не менее двух раз. Это связано с тем, что часть текстов была представлена в виде графических данных, которые были преобразованы в текстовый формат путем полуавтоматического распознавания текста, поэтому в некоторых текстах могли возникнуть ошибки. Таким образом, слова, которые встретились только один раз, зачастую являлись именно словами с ошибками. Итоговый размер словаря составил 143,5 тыс. слов. Фонематические транскрипции создавались автоматически с помощью специально разработанного программного модуля, выполняющего преобразование графема-фонема для поданного на вход списка слов на карельском языке. Более подробно процесс создания фонематического словаря представлен в [35].

Триграммная модель языка обучена с помощью программных средств SRILM [38]. Стоит отметить, что оптимально было бы обучать модель языка на расшифровках спонтанной речи, однако объем таких данных обычно не велик, поэтому чаще всего для этой цели используются письменные тексты. В то же время письменная речь сильно отличается от разговорной, что снижает качество моделей. В ходе исследования триграммная модель языка создавалась двумя способами. При первом способе модель обучалась сразу на всех текстах, включая расшифровки обучающей части речевого корпуса. Второй способ состоял в том, что вначале отдельно обучались две модели языка: одна обучалась на расшифровках обучающей части речевого корпуса, вторая — на остальных текстах. Затем была выполнена линейная интерполяция созданных моделей, при этом коэффициент интерполяции модели, обученной на расшифровках, задавался выше, чем для модели, обученной на письменных текстах. Были проведены эксперименты с использованием разных значений весового коэффициента интерполяции.

Результаты экспериментов по распознаванию карельской речи

Для декодирования речевого сигнала использовался декодер Kaldi на основе взвешенных

конечных преобразователей [8]. Оценка работы системы распознавания речи проводилась по показателю WER. Вначале были проведены эксперименты по автоматическому распознаванию карельской речи с моделью языка, обученной сразу на всех текстовых данных. При использовании акустической модели на основе гауссовых смесей значение WER составило 40,00 %. Результаты экспериментов по автоматическому распознаванию карельской речи с гибридными акустическими моделями на базе архитектуры TDNN с различным временным контекстом и различным отношением числа входов к числу выходов представлены в табл. 3. Наилучшие результаты получены при использовании архитектуры TDNN с пятью скрытыми слоями и временным контекстом [–8, 4] и отношением числа входов к числу выходов, равным 1000/100. Увеличение длины контекста и отношения числа входов/выходов привело к ухудшению качества распознавания, что может быть вызвано переобучением ИНС.

Затем были проведены эксперименты с использованием модели языка, созданной вторым способом, — путем линейной интерполяции модели, обученной на расшифровках аудиоданных, с моделью, обученной на текстах с различными коэффициентами интерполяции. В ходе этих экспериментов использовалась акустическая модель, которая дала наилучшие результаты в предыдущем эксперименте (модель с отношением числа входов к числу выходов, равным 1000/100, и временным контекстом [–8, 4]). Результаты экспериментов представлены в табл. 4, где также указаны значения коэффициента неопределенности (perplexity) для каждой модели языка, вычисленные по текстам из расшифровок тестовой части речевого корпуса. Количество внесловарных слов (out-of-vocabulary words) составило 6 %. Скорость распознавания речи без использования графического процессора на компьютере с многоядерным процессором с тактовой частотой 4 ГГц составила около 1,3 RTF (real-time factor).

Наилучшие результаты были получены при использовании коэффициента интерполяции 0,7 для модели, обученной на расшифровках. Следует отметить, что текстовые данные, полученные из расшифровок, наиболее адекватно отражают разговорный язык, поскольку в них представлена спонтанная речь, на которую не накладываются стилистические правила, характерные для литературного языка.

■ ***Таблица 3.*** Значения WER, полученные с применением гибридных акустических моделей с различными параметрами
■ ***Table 3.*** WER for hybrid acoustic models with different parameters

Длина контекста	Контекст послойно					WER, %, для		
	1	2	3	4	5	500/50[1]	1000/100[1]	2000/200[1]
[–6, 4]	[–1, 1]	{–2, 1}	{–3, 2}	{0}	–	29,46	29,32	31,22
[–6, 6]	[–1, 1]	{–2, 2}	{–3, 3}	{0}	–	31,08	31,49	29,86
[–7, 7]	[–1, 1]	{–2, 2}	{–4, 4}	{0}	–	31,35	31,62	29,05
[–7, 7]	[–2, 2]	{–1, 1}	{–2, 2}	{–2, 2}	{0}	30,14	30,27	30,00
[–8, 4]	[–1, 1]	{–3, 1}	{–4, 2}	{0}	–	30,81	**28,51**	28,65
[–8, 5]	[–2, 2]	{–1, 1}	{–2, 1}	{–3, 1}	{0}	29,46	28,92	28,78
[–8, 8]	[–2, 2]	{–1, 1}	{–2, 2}	{–3, 3}	{0}	29,32	29,32	29,46

[1]Количество входов/выходов.

■ ***Таблица 4.*** Результаты экспериментов с применением различных моделей языка
■ ***Table 4.*** Experimental results with different language models

Способ обучения модели языка	Весовой коэффициент модели языка, обученной на расшифровках	Коэффициент неопределенности	WER, %
На всех текстах сразу	–	4030,06	28,51
Путем линейной интерполяции	0,5	2118,88	26,35
	0,6	2083,11	26,08
	0,7	2095,15	**25,81**
	0,8	2172,63	26,89

Заключение

В работе представлена система автоматического распознавания речи для ливвиковского наречия карельского языка. Для повышения точности работы системы выполнена аугментация речевых данных, кроме того, отдельно обучена модель языка на расшифровках обучающей части речевого корпуса, которая затем интерполировалась с моделью языка, обученной на текстовых данных. Ошибка распознавания слов, полученная в ходе проведенных экспериментов, составила 25,81 %, что, конечно, хуже, чем современные результаты, получаемые для языков с большими ресурсами данных, но находится на уровне мировых результатов для других малоресурсных языков.

Другим важным результатом явилось создание обучающего набора данных, состоящего из звукозаписей на карельском языке, их текстовых транскрипций и собственно текстового корпуса. В представленном исследовании набор данных использован для обучения языковой и акустической моделей, однако его можно использовать и в других исследованиях по карельскому языку в области обработки естественного языка.

Одна из главных проблем, возникших при создании системы автоматического транскрибирования карельского языка, состоит в нехватке данных. Авторы статьи решили эту проблему, во-первых, созданием собственного набора данных и, во-вторых, применением аугментации.

Однако, несмотря на достигнутые результаты, существует потребность в дальнейших исследованиях и улучшении разработанной системы. Так, проблема переключения кодов осталась за рамками настоящего исследования. В общем и целом расширение обучающего набора данных существенно повысит качество разработанной системы. В дальнейшем планируется исследовать метод переноса знаний при обучении акустических моделей, а также применить нейросетевой подход к обучению модели языка.

Результаты настоящего исследования представляют собой вклад в развитие технологий автоматической обработки карельского языка и могут быть использованы при создании подобных систем для иных малоресурсных языков.

Финансовая поддержка

Работа выполнена при финансовой поддержке фонда РНФ (проект № 22-21-00843).

Литература

1. **Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M.** The state and fate of linguistic diversity and inclusion in the NLP world. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6282–6293. doi:10.48550/arXiv.2004.09095
2. **Афанасьева А. А., Муллонен И. И.** Кареловепсский диалог на карте южной Карелии. *Acta Linguistica Petropolitana*, 2020, no. 16(3), с. 9–28. doi:10.30842/alp2306573716301
3. **Krauwer S.** The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proc. of Intern. Workshop on Speech and Computer (SPECOM-2003)*, 2003, pp. 8–15.
4. **Berment V.** *Méthodes pour informatiser des langues et des groupes de langues «peu dotées»*: Doct. Diss. Grenoble, 2004. 278 p. https://theses.hal.science/tel-00006313/document (дата обращения: 28.04.2023).
5. **Cieri Ch., Maxwell M., Strassel S., Tracey J.** Selection criteria for low resource language programs. *Proc. of the Tenth Intern. Conf. on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4543–4549.
6. **Virkkunen A., Rouhe A., Phan N., Kurimo M.** Finnish parliament ASR corpus. *Language Resources and Evaluation*, 2023. https://doi.org/10.1007/s10579-023-09650-7 (дата обращения: 28.04.2023)
7. **Alumäe T., Tilk O., Asadullah.** Advanced rich transcription system for Estonian speech. *Frontiers in Artificial Intelligence and Applications; Ebook*, 2018, vol. 307. doi:10.3233/978-1-61499-912-6-1
8. **Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček P., Qian Y., Schwarz P., Silovský J., Stemmer G., Vesel K.** The Kaldi speech recognition toolkit. *Proc. of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 1–4.
9. **Olev A., Alumäe T.** Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing*, 2022, vol. 10(3), pp. 409–421. doi:10.22364/bjmc.2022.10.3.14
10. **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. 3, pp. 12449–12460. doi:10.48550/arXiv.2006.11477
11. **Марковников Н. М., Кипяткова И. С.** Исследование методов построения моделей кодер-декодер для распознавания русской речи. *Информационно-управляющие системы*, 2019, № 4, с. 45–53. doi:10.31799/1684-8853-2019-4-45-53
12. **Stoian M., Bansal S., Goldwater Sh.** Analyzing ASR pretraining for low-resource speech-to-text translation. *Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2020)*,

2020, pp. 7909–7913. doi:10.1109/ICASSP40776.2020.9053847

13. **Sperber M., Neubig G., Niehues J., Waibel A.** Attention passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 2019, vol. 7, pp. 313–325. doi:10.1162/tacl_a_00270
14. **Salesky E., Sperber M., Waibel A.** Fluent translations from disfluent speech in end-to-end speech translation. *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1 (Long and Short Papers), pp. 2786–2792. doi:10.18653/v1/N19-1285
15. **Godard P., Adda G., Adda-Decker M., Benjumea J., Besacier L., Cooper-Leavitt J., Kouarata G.-N., Lamel L., Maynard H., Mueller M., Rialland A., Stueker S., Yvon F., Zanon-Boito M.** A very low resource language speech corpus for computational language documentation experiments. *Proc. of the Eleventh Intern. Conf. on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 3366–3370.
16. **Bansal S., Kamper H., Livescu K., Lopez A., Goldwater S.** Pre-training on high-resource speech recognition improves low resource speech-to-text translation. *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1 (Long and Short Papers), pp. 58–68. doi:10.18653/v1/N19-1006
17. **Wet de F., Kleynhans N., Compernolle van D., Sahraeian R.** Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems. *South African Journal of Science*, 2017, vol. 113, no. 1–2, pp. 1–9. doi:10.17159/sajs.2017/20160038
18. **Woldemariam Y.** Transfer learning for less-resourced semitic languages speech recognition: The case of Amharic. *Proc. of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 2020, pp. 61–69.
19. **Karunathilaka H., Welgama V., Nadungodage T., Weerasinghe R.** Low-resource Sinhala speech recognition using deep learning. *Proc. of 2020 20th Intern. Conf. on Advances in ICT for Emerging Regions (ICTer)*, 2020, pp. 196–201. doi:10.1109/ICTer51097.2020.9325468
20. **Peddinti V., Povey D., Khudanpur S.** A time delay neural network architecture for efficient modeling of long temporal contexts. *Proc. of INTERSPEECH-2015*, 2015, pp. 3214–3218. doi:10.21437/Interspeech.2015-647
21. **Gupta V., Boulianne G.** Progress in multilingual speech recognition for low resource languages Kurmanji Kurdish, Cree and Inuktut. *Proc. of the 13th Conf. on Language Resources and Evaluation (LREC 2022)*, 2022, pp. 6420–6428.
22. **Biswas A., Menon R., Westhuizen van der E., Niesler Th.** Improved low-resource Somali speech recognition by semi-supervised acoustic and language model training. *Proc. of INTERSPEECH-2019*, 2019. *arXiv preprint*, 2019. arXiv:1907.03064 (дата обращения: 28.04.2023). doi:10.21437/Interspeech.2019-1328
23. **Обухов Д. С.** Разработка современной системы распознавания русскоязычной телефонной речи. *Управление большими системами*, 2021, № 89, с. 106–122. doi:10.25728/ubs.2021.89.4
24. **Kipyatkova I.** Improving Russian LVCSR using deep neural networks for acoustic and language modeling. *Proc. of the 20th Intern. Conf. on Speech and Computer SPECOM-2018*, 2018, vol. 11096, pp. 291–300. doi:10.1007/978-3-319-99579-3_31
25. **Park D. S., Chan W., Zhang Y., Chiu Ch.-Ch., Zoph B., Cubuk E. D., Le Q. V.** Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. of INTERSPEECH-2019*, 2019, pp. 2613–2617. doi:10.21437/Interspeech.2019-2680
26. **Kaneko T., Kameoka H., Tanaka K., Hojo N.** CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion. *Proc. of INTERSPEECH-2020*, 2020, pp. 2017–2021. doi:10.21437/Interspeech.2020-2280
27. **Rebai I., BenAyed Y., Mahdi W., Lorré J. P.** Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, 2017, vol. 112, pp. 316–322. doi:10.1016/j.procs.2017.08.003
28. **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
29. **Kobayashi S.** Contextual augmentation: Data augmentation by words with paradigmatic relations. *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2018, vol. 2 (Short Papers), pp. 452–457. doi:10.18653/v1/N18-2072
30. **Sennrich R., Haddow B., Birch A.** Improving neural machine translation models with monolingual data. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, vol. 1 (Long Papers), pp. 86–96. doi:10.18653/v1/P16-1009
31. **Gokay R., Yalcin H.** Improving low resource Turkish speech recognition with data augmentation and TTS. *Proc. of 2019 16th Intern. Multi-Conf. on Systems, Signals and Devices (SSD)*, 2019, pp. 357–360. doi:10.1109/SSD.2019.8893184
32. **Кипяткова И. С., Кагиров И. А.** Аналитический обзор методов решения проблемы малых наборов данных при создании систем автоматического распознавания речи для малоресурсных языков. *Информатика и автоматизация*, 2022, № 21(4), с. 678–709. doi:10.15622/ia.21.4.2
33. **Ковалева С. В., Родионова А. П.** *Традиционное и новое в лексике и грамматике карельского языка (по данным социолингвистического исследования).* Петрозаводск, КарНЦ РАН, 2011. 138 с.

https://www.booksite.ru/fulltext/koval/text.pdf (дата обращения: 28.04.2023).
34. **Karjalainen H., Ulriikka P., Riho G., Svetlana K.** Karelian in Russia: ELDIA Case-Specific Report, with contributions by Reetta Toivanen, Anneli Sarhimaa and Eva Kühhirt (Studies in European Language Diversity 26). Research consortium ELDIA, 2013. https://phaidra.univie.ac.at/detail/o:314612 (дата обращения: 28.04.2023)
35. **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, № 45(5), с. 82–91. (В печати.)
36. **Saon G., Soltau H., Nahamoo D., Picheny M.** Speaker adaptation of neural network acoustic models using i-vectors. *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 55–59. doi:10.1109/ASRU.2013.6707705
37. **Zhang X., Trmal J., Povey D., Khudanpur S.** Improving deep neural network acoustic models using generalized maxout networks. *Proc. of 2014 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 215–219. doi:10.1109/ICASSP.2014.6853589
38. **Stolcke A., Zheng J., Wang W., Abrash V.** SRILM at sixteen: update and outlook. *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 5–9.

UDC 004.934
doi:10.31799/1684-8853-2023-3-16-25
EDN: YOLUUY

Automatic speech recognition system for Karelian

I. S. Kipyatkova[a], PhD, Tech., Associate Professor, http://orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su
I. A. Kagirov[a], Research Fellow, orcid.org/0000-0003-1196-1117
[a]St. Petersburg Federal Research Center of the RAS, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation

Introduction: There has been a growth in the number of studies devoted to automatic processing of low-resource languages. The lack of training data is a significant obstacle to the development of speech technologies for such languages. **Purpose:** To develop an automatic speech recognition system for Karelian. **Results:** we present a system for automatic speech recognition in Karelian. We have trained acoustic models based on artificial neural networks with time delays and hidden Markov models. We have trained the system with the use of a speech corpus composed of radio broadcast recordings and audio data modified with augmentation techniques. Both written texts and transcripts of a training part of the speech corpus have been involved. We have explored various coefficients to interpolate a language model trained on transcripts with a language model trained on written texts. The best value of the word error rate was 25.81%, which is comparable with the results for other low-resource languages. We have collected a training data set, which includes sound recordings of the Karelian language with transcripts, as well as a text corpus. **Practical relevance:** The results can be of a certain significance for the development of automatic recognition systems not only for Karelian but for other low-resource languages as well. In addition, the developed system may be useful for the researchers of the Karelian language, providing them with an effective tool for recording and processing the Karelian language data.

Keywords — low-resource languages, automatic transcription, the Karelian language, Time Delay Neural Network.

For citation: Kipyatkova I. S., Kagirov I. A. Automatic speech recognition system for Karelian. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2023, no. 3, pp. 16–25 (In Russian). doi:10.31799/1684-8853-2023-3-16-25, EDN: YOLUUY

Financial support
This work was supported financially by the Russian Science Foundation (project No. 22-21-00843).

References

1. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The state and fate of linguistic diversity and inclusion in the NLP world. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6282–6293. doi:10.48550/arXiv.2004.09095
2. Afanasjeva A. A., Mullonen I. I. A Karelian-Veps dialogue on the map of southern Karelia. *Acta Linguistica Petropolitana*, 2020, no. 16(3), pp. 9–28 (In Russian). doi:10.30842/alp2306573716301
3. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proc. of Intern. Workshop on Speech and Computer (SPECOM-2003)*, 2003, pp. 8–15.
4. Berment V. *Méthodes pour informatiser des langues et des groupes de langues «peu dotées»*: Doct. Diss. Grenoble, 2004, 278 p. (In French). Available at: https://theses.hal.science/tel-00006313/document (accessed 28 April 2023).
5. Cieri Ch., Maxwell M., Strassel S., Tracey J. Selection criteria for low resource language programs. *Proc. of the Tenth Intern. Conf. on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4543–4549.
6. Virkkunen A., Rouhe A., Phan N., Kurimo M. Finnish parliament ASR corpus. *Language Resources and Evaluation*, 2023. Available at: https://doi.org/10.1007/s10579-023-09650-7 (accessed 28 April 2023).
7. Alumäe T., Tilk O., Asadullah. Advanced rich transcription system for Estonian speech. *Frontiers in Artificial Intelligence and Applications; Ebook*, 2018, vol. 307. doi:10.3233/978-1-61499-912-6-1
8. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček P., Qian Y., Schwarz P., Silovský J., Stemmer G., Vesel K. The Kaldi speech recognition toolkit. *Proc. of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 1–4.
9. Olev A., Alumäe T. Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing*, 2022, vol. 10(3), pp. 409–421. doi:10.22364/bjmc.2022.10.3.14