УДК 004.934+81'322 **DOI** 10.15622/ia.24.2.9

И.С. Кипяткова, И.А. Кагиров, М.Д. Долгушин

ПРИМЕНЕНИЕ ПРЕДВАРИТЕЛЬНО ОБУЧЕННЫХ МНОГОЯЗЫЧНЫХ МОДЕЛЕЙ ДЛЯ РАСПОЗНАВАНИЯ КАРЕЛЬСКОЙ РЕЧИ

Кипяткова И.С., Кагиров И.А., Долгушин М.Д. **Применение предварительно обученных многоязычных моделей для распознавания карельской речи.**

Аннотация. В настоящей статье описывается экспериментальное исследование, направленное на решение проблемы обучения моделей для распознавания речи в условиях малого объема обучающих речевых и текстовых данных. Подробно рассматриваются существующие подходы к решению данной проблемы, в частности, использование преодобученных многоязычных моделей и аугментация данных. В работе проведена адаптация многоязычных моделей на базе Wav2Vec и Whisper к ливвиковскому наречию карельского языка и проведено исследование применения внешней языковой модели для повышения точности распознавания интегральной системы. Кроме того, в статье описаны специально собранная и подготовленная речевая база данных и базовая система распознавания, созданная на основе тулкита Kaldi. Приведены количественные результаты тестирования, которые подтверждают эффективность выбранных методов: так, использование моделей на архитектуре Трансформер, в частности, Wav2Vec, позволило достичь более высоких показателей, чем у базовых моделей, обученных с помощью программных средств Kaldi. Дообучение моделей Wav2Vec снизило количество неправильно распознанных слов до 24,73% на валидационной и до 25,25% на тестовой выборках, а использование модели Wav2Vec-BERT 2.0 с внешней языковой моделью дополнительно уменьшило количество неправильно распознанных слов до 17,12% и 17,72% соответственно. Статья адресована, в первую очередь, специалистам, занимающимся разработкой систем автоматического распознавания речи для малоресурсных языков и распознаванием речи на прибалтийско-финских языках, в частности, результаты этой работы могут найти практическое применение в полевых исследованиях, при записи текстов на карельском.

Ключевые слова: малоресурсные языки, карельский язык, переключение кодов, предобученные модели, машинное обучение, речевой корпус.

1. Введение. В последние годы наблюдается рост интереса к исследованиям в области автоматического распознавания речи для малоресурсных языков. Это связано, среди прочего, с актуальностью проблемы исчезновения миноритарных языков коренных народов. Существенным препятствием, с которым сталкиваются разработчики подобных систем, является дефицит данных для обучения акустических и языковых моделей.

За последние годы технологии автоматического распознавания речи претерпели значительные изменения, перейдя от классических модульных систем к интегральным. В традиционных системах выделение признаков, акустическое и языковое моделирование, а также декодирование выполняются отдельными модулями. Интегральный (или end-to-end) подход, наоборот, объединяет все эти

модули в одну глубокую нейронную сеть, которая обучается как единое целое. Интегральные системы превосходят традиционные по точности и скорости распознавания, однако одним из ключевых требований для обучения интегральной системы является доступность больших объемов данных. Малоресурсные языки по определению не обладают оцифрованными данными большого объема, что требует разработки особых подходов для применения современных методов машинного обучения к их материалу.

Дополнительной сложностью при создании систем распознавания речи для малоресурсных языков является феномен переключения кодов (англ. code-switching) – перехода с одного языка на другой в процессе речи. Поскольку малоресурсные языки часто существуют в условиях полиязычной среды, переключение кодов особенно характерно именно для них. Это явление дополнительно обостряет проблему недостатка обучающих данных, поскольку переключение кодов снижает однородность и согласованность речевых корпусов.

В настоящей статье представлено продолжение серии экспериментов по автоматическому распознаванию речи на малоресурсном карельском языке [1 – 4]. На примере карельского языка описаны этапы создания подобной системы, включая подготовку речевых данных, обучение акустических и языковых моделей. Приведены количественные результаты тестирования разработанных моделей, подтверждающие эффективность применённых методов.

2. Основные методы решения проблемы нехватки данных при обучении систем распознавания речи. Как следует из предыдущего раздела, «прямое» применение интегрального подхода к материалу малоресурсных языков оказывается проблематичным, и среди важнейших задач, которые должны быть решены до этапа обучения системы, оказывается создание набора обучающих данных достаточного объема.

Одним из способов расширения объема обучающих данных является аугментация. Среди основных методов аугментации аудиоданных можно перечислить [5]:

- изменение высоты голоса, темпа речи, громкости речи;
- модификация речевых признаков (например, добавление случайных значений к речевым признакам);
- изменение спектрограммы;
- преобразование голоса (англ. voice conversion);
- синтез речи.

Аугментация текстовых данных может выполняться следующими способами:

- использование данных другой предметной области;
- использование машинного перевода;
- модификация текста путем случайной замены/вставки/удаления слов или символов;
- генерация текста с помощью искусственных нейронных сетей (ИНС).

Другим подходом к решению проблемы нехватки данных, который показал свою эффективность при обучении нейронных сетей, является метод переноса знаний (англ. transfer learning). Суть этого метода состоит в предварительном обучении модели на большом объеме нецелевых данных с последующим дообучением модели на малом объеме целевых данных. Таким образом, основной принцип переноса знаний – это применение знаний нейросети, обученной на одной задаче, к другой задаче с предварительным дообучением.

В настоящее время существуют предварительно обученные многоязычные модели для распознавания речи, которые находятся в открытом доступе. Одной из таких моделей является модель Wav2Vec 2.0 [6], использующая метод самообучения, подразумевающий начальное обучение речевых представлений на большом объеме неразмеченных данных, а затем модель дообучается на меньшем объеме размеченных данных. Модель Wav2Vec 2.0 состоит из трех блоков: блок извлечения признаков, в качестве которого используется многослойная сверточная нейронная сеть, трансформерный кодер и модуль квантования, который принимает на вход все различные представления речевого сигнала, сгенерированные блоком извлечения признаков, и сводит их к конечному набору речевых единиц.

Существуют различные версии модели, например, Base и Large, различающиеся количеством обучаемых параметров, а также многоязычная вариация модели, использующая кросс-языковые представления, XLS-R (англ. cross-lingual speech representations), которая может быть использована при создании моноязычных моделей [7, 8]. В работе [9] эта модель, под названием MMS (Massively Multilingual Speech), была адаптирована для материала более чем 1000 языков, одним из которых был карельский. Авторам упомянутой работы удалось добиться хорошего уровня распознавания за счет эффективного использования слоев Адаптеров [1] и постепенного добавления языков на этапе обучения [10]. Кроме того, как было

[1] https://huggingface.co/docs/peft/conceptual_guides/adapter

показано в работе [11], объединение языковых моделей и лингвистических кодеров, например, BERT [12], с акустическим кодером может быть перспективным в задачах распознавания речи с ограниченными ресурсами. Некоторые из этих моделей, такие как WavLM [1], использующая этап кластеризации для получения выравненных BERT-подобных потерь, а также шумоподавление и управляемое смещение относительной позиции кодов, и W2V2-BERT V2, развивающая архитектуру Wav2Vec за счет замены механизма внимания на модель Conformer с применением каузального глубинного слоя, могут демонстрировать результаты, соответствующие современному техническому уровню [13, 14] распознавания речи.

Другой широко распространенной предобученной моделью является модель Whisper от компании OpenAI, обученная на 680 тыс. часах размеченных многоязычных данных. Whisper основана на архитектуре Трансформер, которая состоит из кодера и декодера. Кодер состоит из двух сверточных слоев, за которыми следует синусоидальное позиционное кодирование и блоки Transformer. Декодер использует обученные векторы позиционного кодирования и содержит то же количество блоков Трансформер, что и кодер. Архитектура Whisper подробно описана в [15].

Существует несколько реализаций модели Whisper – Tiny, Base, Small, Medium, Large, – которые отличаются количеством используемых параметров. Так же, как и Wav2Vec, модель Whisper может быть дообучена на целевых данных, однако, как было показано в нескольких исследованиях, дообучение модели Whisper показывает результаты распознавания хуже, нежели Wav2Vec. Например, в работе [16] было проведено исследование применения моделей Wav2Vec 2.0 и Whisper для распознавания мальтийской речи с переключением кода на английский. Авторы выполнили дообучение моделей Wav2Vec XLS-R с 300 млн и с 3 млрд параметров, в также были проведены эксперименты с моделями Whisper Tiny, Small и Large на речевых данных различного объема от 10 минут до 100 часов. Наилучшие результаты показала модель XLS-R 2B с 2 млрд параметров при дообучении на 50 часах речевых данных, при этом значение WER (количество неправильно распознанных слов, англ. word error rate) составило 8,53%, значение CER (количество неправильно распознанных символов, англ. character error rate) – 1,93% на тестовой части корпуса CommonVoice и 24,98% – WER, 8,37% – CER на корпусе MASRI. Сравнимые результаты были получены при

[1] https://huggingface.co/docs/transformers/model_doc/wavlm

дообучении модели на данных объемом примерно 10 часов. Модель Whisper показала существенно худшие результаты, и при использовании модели Whisper Tiny значение WER во всех экспериментах составило 100%.

Другое исследование, посвященное сравнению моделей Wav2Vec 2.0 и Whisper, представлено в статье [17]. Суть этой работы состояла в дообучении моделей Wav2Vec 2.0 Base и XLSR-53 вместе с моделями Whisper Small и Large на корпусе казахской речи. Модели Wav2Vec 2.0 превзошли модели Whisper по точности распознавания. Наилучшие результаты были получены для модели Wav2Vec 2.0 Base (WER=9,8%, CER=2,7%), при этом с применением многоязычной модели Whisper Large значение WER составило 19,8%, CER – 4,1%.

Еще одной проблемой, которая часто возникает при обработке малоресурсных языков, является проблема переключения кодов. Существуют два основных подхода к решению данной проблемы. Первый предполагает определение границ разноязычных фрагментов речи и их обработку моноязычной системой. Для определения языка используют акустические признаки (например, i-вектора или bottleneck-признаки [18]), лексические данные (теги частей речи [19], триггерные слова [20]) или их комбинацию [21]. Второй подход использует многоязычные системы распознавания речи, требующие унификации алфавитов и фонем [22].

Стоит отметить, что зачастую особые сложности вызывает нехватка текстовых данных для языкового моделирования, так как переключение кодов реже встречается в письменной речи [23, 24]. Для аугментации данных применяют частичный автоматический перевод [25] и генерацию текста с помощью ИНС [26]. Среди других методов можно перечислить использование класс-ориентированных моделей [27], факторных моделей языка с тегами переключения [28] и двуязычных векторных представлений слов, основанных на параллельных корпусах [29].

В последнее время все большее распространение приобретает применение предварительно обученных многоязычных моделей, таких как mBERT (обученная на текстах из Википедии на 104 языках) [12] или XLM RoBERTa (обученная на 2,5 ТБ отфильтрованных данных CommonCraw на 100 языках) [30]. Так, в работе [31] исследуется применение больших предобученных моделей (GPT-2 и BERT) для распознавания речи для ряда африканских языков в условиях переключения кодов с английским и делается вывод о том, что применение подобных моделей эффективно позволяет снизить WER. Также в данной работе делается вывод о перспективности методики

формирования обучающего корпуса путем чередования текстовых данных из моноязычных корпусов разных языков.

Еще одним способом повышения точности распознавания речи является использование внешней языковой модели [32, 33], которая улучшает синтаксическую и семантическую интерпретацию текста за счет более точного вероятностного представления целевого языка, при этом могут использоваться как статистические, так и нейросетевые модели. Несмотря на то, что такая модель представляет собой отдельный модуль, чаще всего подобные системы называют интегральными системами с внешней моделью языка [34]. В последующих разделах будут продемонстрированы практические решения задачи распознавания речи в рамках интегрального подхода на материале ливвиковского наречия малоресурсного карельского языка.

3. Методика исследования. В рамках данного исследования был проведен сравнительный анализ применения предобученных моделей на основе Wav2Vec и Whisper с базовой системой на основе Kaldi [35] для задачи распознавания карельской речи. Для обучения и тестирования моделей использовалась база данных AnKaS[1] – «База данных аннотаций речевых записей на карельском языке (AnKaS – Database of Annotations of Karelian Speech Recordings)» (далее – БД AnKaS), описание которой приведено ниже в разделе 4.

В качестве базовой системы использовалась многомодульная система распознавания карельской речи на основе Kaldi, включающая гибридную акустическую модель, объединяющую ИНС и скрытые марковские модели (СММ) – ИНС/СММ, а также модели языка двух типов: статистическую (используемую на этапе декодирования) и нейросетевую (применяемую на этапе переоценки списка гипотез распознавания). Более подробно базовая система распознавания описана в разделе 5.

Выбраны предварительно обученные интегральные модели на основе Wav2Vec и Whisper, и выполнено их дообучение. Обоснование выбора предобученных моделей и процесс их дообучения представлены в разделе 6.

Проведены эксперименты по распознаванию карельской речи с использованием многомодульной и интегральных моделей, а также эксперименты с применением внешней языковой модели дополнительно к интегральной. Оценка качества распознавания речи осуществлялась по показателю WER, описание экспериментов и полученных результатов приведено в разделе 7.

[1] https://github.com/IrinaKipyatkova/AnKaS

4. База данных аннотаций речевых записей на карельском языке. БД AnKaS представляет собой набор аннотаций записей карельской речи (ливвиковское наречие) из 13 радиопередач программы «Kodirandaine» («Родной берег»). Оригинальные аудиофайлы находятся в открытом доступе на сайте ГТРК «Карелия»[1]. Была выполнена расшифровка аудиозаписей. В БД AnKaS были включены аннотации только таких речевых отрезков, в которых отсутствует фоновый шум, речевые сбои и одновременная речь нескольких дикторов. Всего были созданы аннотации для 4385 фраз (4,5 часа речи). Основные характеристики БД AnKaS представлены в таблице 1.

Таблица 1. Характеристики БД AnKaS

Параметр	**Значение**
Общее количество дикторов	17 (7 мужчин, 10 женщин)
Длительность аннотированных речевых данных	4,5 часа
Количество фраз	4385
Количество словоупотреблений	32037
Количество уникальных слов	9117

Логическая структура БД AnKaS представлена на рисунке 1. БД представлена в формате JSON. Аннотация речевых записей каждого диктора содержится в отдельном .json файле, при этом используются следующие ключи:

- “phrase_id” – номер фразы у данного диктора;
- “link”– интернет-ссылка на аудиозапись на сайте ГТРК «Карелия»;
- “time_start” – время начала фразы;
- “time_end” – время окончания фразы;
- “sentence” – текстовая расшифровка;
- “sentence_rus” – текстовая расшифровка с метками переключения языка на русский, при этом русскоязычный текст заключается в треугольные скобки и помечается тегом «rus».

[1] https://tv-karelia.ru/kodirandaine-rodnoy-bereg/

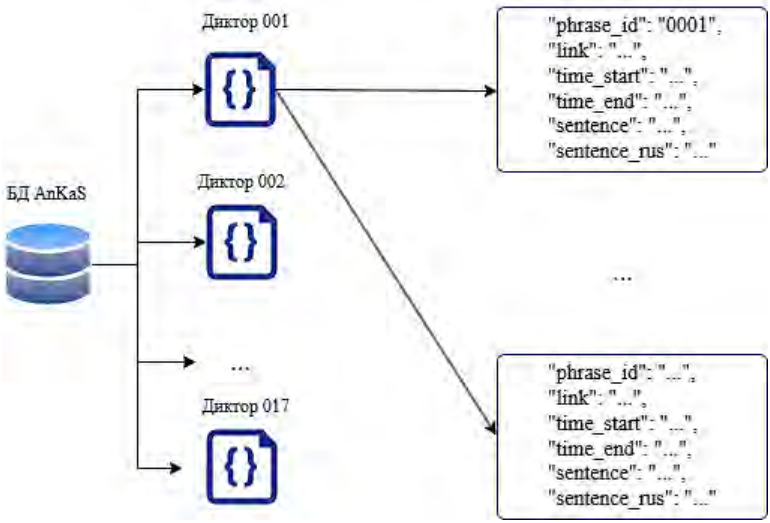

Рис. 1. Логическая структура БД AnKaS

Пример аннотации для одной фразы показан на рисунке 2.


```
{
    "phrase_id": "0002",
    "link": "https://tv-karelia.ru/wp-content/uploads/2022/01/Kirvesmies-Aleksandr-Ivanov.mp3",
    "time_start": "114.7394720",
    "time_end": "120.5044720",
    "sentence": "vot d'ad'a miša jakovlev täs susiedu d'ad'a pet'a jakovlev täs test'u opasti",
    "sentence_rus": "<vot d'ad'a miša jakovlev>rus täs susiedu <d'ad'a pet'a jakovlev>rus täs test'u opasti"
},
```


Рис. 2. Пример аннотации одной фразы

Особо стоить отметить, что в аннотации содержится в том числе и информация о переключении кода с карельского на русский. Случаи переключения кодов, обнаруженные в собранных данных, немногочисленны со статистической точки зрения (только 1,13% лексических единиц были помечены тегами переключения кодов). Однако, принимая во внимание тот факт, что единственным источником языковых данных были радиопередачи (то есть в записях присутствуют не спонтанные нарративы, а подготовленная речь, звучащая в формальной обстановке), это число всё равно значимо.

Собранный корпус использовался для обучения и тестирования системы распознавания карельской речи, при этом данные были разделены на обучающую, валидационную и тестовую выборки. 80% высказываний были использованы для обучения акустической модели, 10% – для валидации и настройки гиперпараметров, а 10% – для финального тестирования. Аугментация обучающей части корпуса выполнялась путем совместного и поочередного изменения частоты основного тона (ЧОТ) и темпа речи. Модификация ЧОТ выполнялась на количество полутонов, полученных случайным образом из равномерного распределения в диапазоне [-2, 2], а изменение темпа речи осуществлялось с помощью коэффициента, случайно выбранного из равномерного распределения в диапазоне [0,7, 1,3]. Для выполнения описанных модификаций речевого сигнала использовался инструментарий SoX.[1] В результате размер обучающей выборки был увеличен до 13,5 часов. Более подробно процесс подготовки речевого материала описан в работе [36].

5. Система распознавания карельской речи на основе Kaldi. Базовая система распознавания карельской речи была создана с помощью тулкита Kaldi [35]. Для акустического моделирования использовалась гибридная СММ/ИНС модель [2] на основе факторизованной ИНС с временными задержками (англ. TDNN-F), то есть ИНС с временными задержками (TDNN), в которой размерность слоев сокращена путем сингулярного разложения [37]. Архитектура ИНС показана на рисунке 3.

Архитектура сети состоит из трех блоков TDNN-F. Первый блок имеет три слоя TDNN-F, обрабатывающих входные вектора во временном интервале {-1,0,1}. Второй блок содержит один слой TDNN-F без объединения с временными шагами (как предыдущим, так и последующим). Третий блок состоит из 10 слоев TDNN-F, обрабатывающих временные шаги {-3,0,3}. Размер слоя TDNN-F равен 1024. В качестве активационной функции использовалась функция ReLU. В модели применялся метод пропуска соединений, аналогичный представленному в [37], при этом выходные данные каждого слоя (кроме первого слоя) добавляются к выходным данным предыдущих слоев с коэффициентом 0,66. За слоями TDNN-F располагается линейный слой размером 256. Обучение осуществлялось в течение 8 эпох, при этом коэффициент скорости обучения уменьшался со значения 0,0005 до 0,00005. Число эпох обучения, коэффициент скорости обучения и другие гиперпараметры моделей были выбраны эмпирически. Входными данными для

[1] http://sox.sourceforge.net/sox.html

нейронной сети были мел-частотные кепстральные коэффициенты (англ. MFCC), при этом для адаптации к речи диктора к ним был добавлен 100-мерный i-вектор [38].

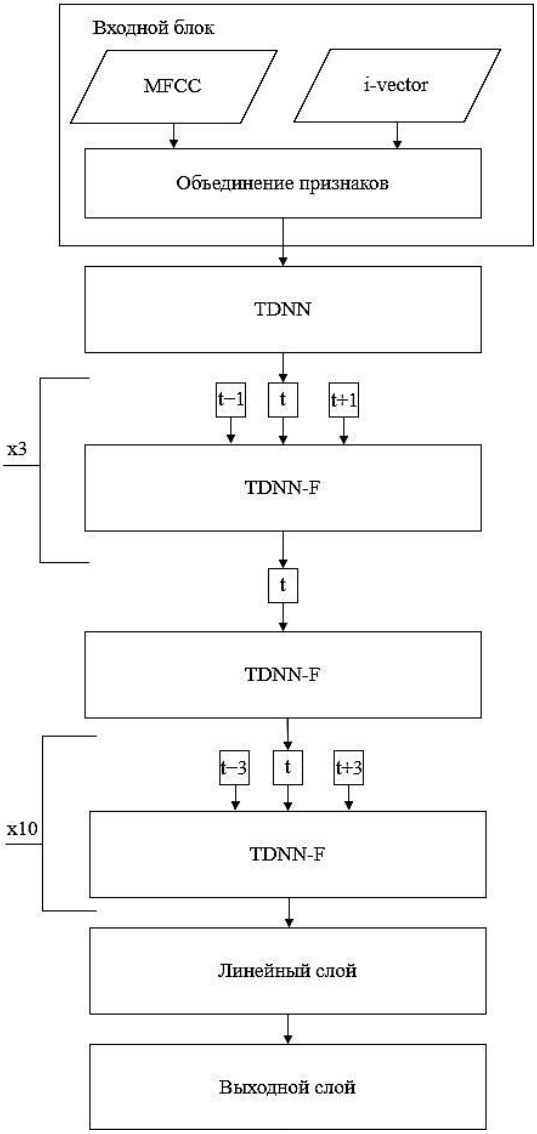

Рис. 3. Архитектура сети TNDNN-F

Для языкового моделирования были созданы два типа моделей: статистическая, которая использовалась на этапе декодирования, и нейросетевая, которая применялась для переоценки списка гипотез распознавания (N-best list). Статистическая модель представляла собой модель на основе триграмм слов, и она была обучена с помощью инструментария SRI Language Modeling Toolkit (SRILM) [39].

Для обучения нейросетевой модели был использован инструментарий TheanoLM [40]. Нейросетевая модель состояла из проекционного слоя размером 500 и двух слоёв LSTM размером 512. В качестве метода оптимизации использовался метод моментов Нестерова. Размер батча был равен 16. Была выполнена линейная интерполяция нейросетевой модели с триграммной, при этом коэффициент интерполяции для нейросетевой модели был равен 0,6. Для обучения языковых моделей использовались текстовые данные, полученные из открытого корпуса вепсского и карельского языков «VepKar» [1] и публикаций на карельском языке, а также из расшифровок обучающей части речевого корпуса. Подробно процесс обработки текстовых данных и создания моделей языка описан в [2, 36].

Текстовый корпус использовался для формирования словаря системы, в него вошли все слова из расшифровок обучающей части корпуса и слова из других текстовых материалов, которые встречались в нём не менее двух раз. Объём словаря системы составил 143907 слов.

Фонематические транскрипции для словаря создавались автоматически. В карельском языке фиксированное ударение, которое всегда падает на первый слог, а гласные не подвержены редукции. Важной особенностью карельского языка является наличие смыслоразличительной долготы как для гласных, так и для согласных. В ходе предыдущего исследования [3] было определено, что наилучшее результаты распознавания речи дает использование такого фонемного инвентаря, в котором долгие гласные являются отдельными фонемами, а для слов с долгими согласными было создано по две альтернативные транскрипции: с интерпретацией долгих согласных как удвоенных кратких и без удвоения фонем. Такое решение необходимо для учета возможной редукции долгого согласного в разговорной речи. В результате процесс автоматического создания транскрипции включает в себя определение ударения, обработку удвоенных фонов и определение палатализованных согласных (перед гласными переднего ряда). Из-за присутствия в расшифровках речевого корпуса русскоязычных слов, словарь

[1] http://dictorpus.krc.karelia.ru/ru

транскрипций включал в себя и русские слова, транскрипции для которых также создавались автоматически. Более подробно проблемы русской фонемной транскрипции обсуждаются в предыдущих работах [41].

6. Выбор предобученной модели для создания системы распознавания карельской речи. Проведен анализ готовых предварительно обученных моделей для распознавания карельской речи, в результате которого было принято решение выбрать следующие интегральные модели, основанные на архитектурах Wav2Vec и Whisper:

- Wav2Vec2.0 Large-Uralic-VoxPopuliV2;
- WavLM Large;
- MMS 1BAll;
- W2V2-BERT 2.0;
- Whisper Small;
- Whisper Medium;
- Whisper Large V2 Distilled.

Данные модели были выбраны по той причине, что они показывают одни из лучших результатов при многоязычном распознавании речи и часто изучаются в контексте применения к малоресурсным языкам [42, 43]. Модель MMS 1B All [9] была рассмотрена, в том числе, еще и потому, что ее авторами заявляется поддержка распознавания как собственно-карельского, так и ливвиковского наречий карельского языка.

В таблице 2 представлено краткое описание этих моделей и параметров, использовавшихся при дополнительном обучении. Дообучение моделей проводилось с использованием фреймворка Transformers [44].

Возможность использования оригинальных моделей Whisper Large v2, Whisper Large v3 и Whisper Large v3 Turbo, показывающих одни из лучших результатов распознавания для английского и русского, не была изучена в рамках настоящей работы по причине высоких требований к вычислительным ресурсам, налагаемых перечисленными моделями.

Важно отметить, что дообучение всех моделей происходило в течение 10000 шагов. Данное число было определено эмпирически. При большем количестве шагов снижение WER на валидационной выборке и функции потерь на обучающей выборке переставало происходить, поэтому данное количество шагов при обучении всех моделей оказалось оптимальным.

Размеры батчей и число шагов аккумуляции градиента подбирались индивидуально в силу разных требований к вычислительным ресурсам, необходимым для дополнительного обучения данных моделей. При возможности предпочтение отдавалось большему размеру батча, но авторы не исключают возможности дополнительного улучшения результатов при увеличении размеров батчей.

Таблица 2. Описание моделей на базе Wav2Vec и Whisper

Название	Кол-во параметров	Кол-во данных (часы), разметка	Число языков*	ЯМ**	Параметры дообучения
Wav2Vec2.0 Large Uralic VoxPopuli V2[1]	300 млн	42,5 тыс., неразмеченные	3	нет	10 тыс. шагов обучения, размер батча – 8, шагов накопления градиента – 4
WavLM Large[2]	316 млн	94 тыс., неразмеченные	1***	нет	10 тыс. шагов обучения, размер батча – 4, шагов накопления градиента – 8
MMS 1B All[3]	1000 млн	49 тыс., размеченные	1162	нет	10 тыс. шагов обучения, размер батча – 8, 4 шагов накопления градиента
W2V2-BERT 2.0[4]	600 млн	4,5 млн, неразмеченные	143	нет	10 тыс. шагов обучения, размер батча – 2, шагов накопления градиента – 16
Whisper Small[5]	244 млн	680 тыс., размеченные	99	да	10 тыс. шагов обучения, размер батча – 16, шагов накопления градиента – 2
Whisper Medium[6]	769 млн	680 тыс., размеченные	99	да	10 тыс. шагов обучения, размер батча – 16, шагов накопления градиента – 2
Whisper Large V2 Distilled[7]	756 млн	22 тыс. размеченные	1***	да	10 тыс. шагов обучения, размер батча – 16, шагов накопления градиента – 2

* Все модели, кроме WavLM Large и Whisper Large V2 Distilled, были обучены в том числе и на уральских языках. Wav2Vec2.0 Large Uralic VoxPopuli V2 обучен исключительно на уральских языках. ** Языковая модель. *** английский.

В моделях на основе Wav2Vec и Whisper при токенизации не учитывалась длительность звуков, в отличие от системы, созданной на базе Kaldi. В ходе предварительных экспериментов исследовалась интерпретация долгих звуков как отдельных фонем, однако она не оказала существенного влияния на результаты. Кроме того,

[1] https://huggingface.co/facebook/wav2vec2-large-uralic-voxpopuli-v2
[2] https://huggingface.co/microsoft/wavlm-large
[3] https://huggingface.co/facebook/mms-1b-all#model-details
[4] https://huggingface.co/docs/transformers/model_doc/wav2vec2-bert
[5] https://huggingface.co/openai/whisper-small
[6] https://huggingface.co/openai/whisper-medium
[7] https://huggingface.co/distil-whisper/distil-large-v2

в работе [43], описывающей интегральную систему для финского, также не учитывалась длительность гласных, хотя в финском языке долгота является смыслоразличительным признаком фонем.

Поскольку модель Whisper не обладает встроенной поддержкой ливвиковского наречия карельского языка, было принято решение дообучить модель для финского на базе Whisper. Возможность добавления нового языка к уже имеющимся в Whisper существует, однако она сопряжена с необходимостью полного переобучения токенизатора, основанного на модели GPT-2, что является неоптимальным решением, связанным с повышением ресурсозатратности. Впрочем, подобный подход был успешно применен в работе [43], где веса для финского языка дообучались на данных по северному диалекту саамского.

Как было замечено выше (раздел 2), добавление языковой модели к предобученной может значительно повысить точность распознавания, в особенности для материала малоресурсных языков [33]. Поэтому были проведены предварительные исследования по интеграции разработанной триграммной языковой модели к некоторым из предобученных моделей, в частности, к Wav2Vec2.0 Large Uralic VoxPopuli V2, MMS-1B All и W2V2 BERT 2.0. Результаты этих экспериментов представлены в следующем разделе.

7. Результаты экспериментов по распознаванию речи. Результаты экспериментов по распознаванию речи на карельском языке с использованием различных типов акустических моделей, полученные на валидационной (Dev) и тестовой (Test) частях корпуса, представлены в таблице 3. В качестве метрики был принят WER.

Вначале были проведены эксперименты по распознаванию речи с использованием модели TDNN-F/CMM, обученной на речевых данных без аугментации. Значение WER, полученное при использовании модели TDNN-F для распознавания валидационной части корпуса, оказалось равно 28,96%, а для тестовой части 30,58%. Использование модели, обученной на аугментированных данных, позволило снизить значение WER до 27,13% на валидационной части корпуса и до 28,77% – на тестовой, таким образом, относительное снижение WER составило 6%. В то же время, выполнение переоценки списка лучших 500 гипотез распознавания с помощью нейросетевой модели языка, интерполированной с триграммной, позволило снизить WER до 25,44% и 27,20% на валидационном и тестовом корпусах соответственно.

Таблица 3. Результаты экспериментов по распознаванию карельской речи

Акустическая модель	**Языковая модель**	**Значение WER, %**	
		Корпус Dev	Корпус Test
TDNN-F/CMM (без аугментации)	Триграммная	28,96	30,58
TDNN-F/CMM	Триграммная	27,13	28,77
TDNN-F/CMM	Триграммная + нейросетевая	25,44	27,20
Wav2Vec2.0 Large Uralic VoxPopuli V2	–	24,73	25,25
Wav2Vec2.0 Large Uralic VoxPopuli V2	Триграммная	19,69	19,83
WavLM Large	–	34,74	38,34
WavLM Large	Триграммная	26,72	29,04
MMS 1B All	–	31,56	32,06
MMS 1B All	Триграммная	24,29	24,99
W2V2-BERT 2.0	–	18,84	20,39
W2V2-BERT 2.0	Триграммная	**17,39**	**17,86**
Whisper Small	–	32,22	35,25
Whisper Medium	–	25,54	28,54
Whisper Large V2 Distilled	–	28,38	30,75

Использование дообученной модели WavLM Large с внешней языковой показало результаты, сравнимые с результатами, полученными с применением моделей TDNN-F/CMM, при этом значения WER составили 26,72% и 29,04% на валидационном и тестовом корпусах соответственно. Однако без языковой модели WavLM Large показала наихудшие результаты. Возможно такое отставание в результатах от других моделей может быть объяснено тем, что модель была предобучена только на английском языке и заведомо не предполагалась быть многоязычной.

Дообучение моделей MMS также показало результаты, сравнимые с базовой моделью, несмотря на то что модели MMS уже имели веса для карельского языка. Было получено значение WER, равное 31,56% на валидационном корпусе и 32,06% на тестовом. Применение триграммной модели значительно улучшило результаты, что привело к достижению 24,29% WER на валидационных данных и 24,99% на тестовых.

Сопоставимые результаты показала модель Wav2Vec, предобученная на уральских языках. Использование данной модели позволило снизить WER до 24,73% на валидационном корпусе и до 25,25% на тестовом. Применение внешней языковой модели привело

к дальнейшему снижению WER до 19,69% на валидационном корпусе и до 19,83% на тестовом.

Наилучшие результаты были получены с помощью модели W2V2-BERT 2.0. Без использования триграммной модели значение WER составило 18,84% на валидационном корпусе и 20,39% – на тестовом. Добавление внешней языковой модели позволило дополнительно улучшить результаты. При этом значение WER составило 17,39% на валидационном корпусе и 17,86% – на тестовом.

Модели на базе Whisper показали результаты хуже, чем Wav2Vec-модели. При этом наилучшие результаты были получены с моделью Whisper Medium; значение WER составило 25,54% и 28,54% на валидационных и тестовых данных соответственно. Возможно, что лучшие результаты могли быть получены с модель Whisper Large v3, однако высокие требования к вычислительным ресурсам для дообучения этой модели не позволили проверить это предположение. Также авторы предполагают, что сравнительно высокая ошибка при использовании моделей на основе Whisper может быть связана с разницей в длительности записей, использовавшихся при обучении и дообучении. Аудиозаписи в использованном наборе не превышают по длительности 10 секунд, тогда как Whisper обучался на записях длительностью в 30 секунд, из-за чего в целях нормирования заполняется нулями. В работе [45] отмечалось позитивное влияние на дообучение Whisper от использования более длинных и шумных записей, при большей устойчивости к шумам по сравнению с Wav2Vec. В контексте карельского языка данную гипотезу только предстоит изучить в будущем.

Далее на валидационной части была проведена настройка гиперпараметров языковой модели:

– alpha – весовой коэффициент, регулирующий значимость модели языка;

– beta – весовой коэффициент, регулирующий длину выходной последовательности; чем больше значение beta, тем меньше длина выходной последовательности.

Значения WER, полученные с использованием модели Wav2Vec2-BERT 2.0, на валидационной части корпуса при различных параметрах alpha и beta языковой модели представлены на рисунке 4. На рисунке 4 более темным цветом выделена область с лучшими результирующими WER при соответствующих alpha и beta. Затем языковая модель с параметром alpha, равном 0,8, и различных значениях beta, продемонстрировавших наилучшие результаты на

валидационной части, была проверена на тестовой части набора, полученные значение представлены в таблице 4.

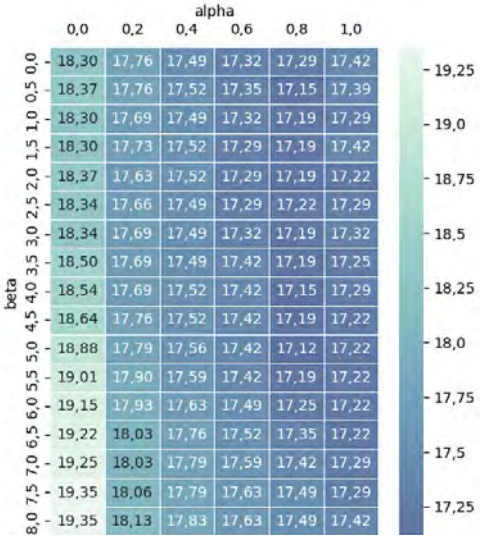

Рис. 4. Значения WER на валидационной части корпуса при различных параметрах alpha и beta языковой модели

Таблица 4. Результаты экспериментов при различных значениях параметра beta и значении alpha, равном 0,8

beta	**Значение WER, %**	
	Корпус Dev	Корпус Test
0,5	17,15	17,79
2,0	17,19	17,79
4,0	17,15	17,75
4,5	17,19	**17,72**
5,0	**17,12**	**17,72**

Наилучший результат на обеих выборках был получен при использовании alpha и beta равных 0,8 и 5,0 соответственно и составил 17,12 и 17,72 на валидационной и тестовой частях, тем самым позволив дополнительно улучшить результат, показанный моделью W2V2-BERT 2.0, на 0,14% на тестовой части. Интересно отметить, что низкое значение ошибки на тестовой части было получено также при beta 4,5 и 4,0, хотя это отличается от результатов, продемонстрированных на валидационной части.

Таким образом, наилучшие результаты были получены при использовании дообученной модели W2V2-BERT 2.0 и внешней языковой модели, что подтвердило эффективность данной архитектуры в задачах малоресурсного распознавания речи.

8. Заключение. В настоящей статье были рассмотрены основные методы, применяемые для решения проблем, связанных с созданием систем распознавания речи для малоресурсных языков в условиях недостаточного объема обучающих данных и переключения кодов. Приведено описание БД AnKAS, содержащей аннотации речевых данных на карельском языке, а также описаны модели распознавания речи для малоресурсного карельского языка, которые создавались с применением двух основных подходов традиционного (многомодульного) и интегрального с использованием предварительно обученных многоязычных моделей.

Результаты экспериментов показывают эффективность созданных моделей, полученные значения WER, равные 17,12% на валидационной выборке и 17,72% на тестовой выборке, соответствуют мировым результатам для малоресурсных языков. Однако стоит отметить, что несмотря на то, что для разговорной карельской речи характерно переключение кода с карельского на русский, в обучающих речевых данных таких явлений было немного. Именно поэтому коллективом авторов настоящей статьи ведется работа по подготовке корпуса с образцами разговорной карельской речи с переключением кодов. Дальнейшая работа будет посвящена обучению акустических и языковых моделей карельского языка с поддержкой переключения кода с карельского на русский.

Литература

1. Кипяткова И.С., Кагиров И.А. Система автоматического распознавания карельской речи // Информационно-управляющие системы. 2023. № 3. С. 16–25.
2. Kipyatkova I., Kagirov I. Deep Models for Low-Resourced Speech Recognition: Livvi-Karelian Case // Mathematics. 2023. vol. 11. no. 18. DOI: 10.3390/math11183814.
3. Kipyatkova I., Kagirov I. Phone Durations Modeling for Livvi-Karelian ASR // Proceedings 25th International Conference Speech and Computer (SPECOM 2023). Springer LNCS. 2023. vol. 14339. pp. 87–99. DOI: 10.1007/978-3-031-48312-7_7.
4. Kipyatkova I., Kagirov I., Dolgushin M., Rodionova A. Towards a Livvi-Karelian End-to-End ASR System // Proceedings 26th International Conference on Speech and Computer (SPECOM 2024). 2024. vol. 15299. pp. 57–68. DOI: 10.1007/978-3-031-77961-9_4.
5. 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. Т. 21. № 4. С. 678–709. DOI: 10.15622/ia.21.4.2.

6. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Advances in Neural Information Processing Systems. 2020. vol. 33. pp. 12449–12460.
7. Conneau A., Baevski A., Collobert R., Mohamed A., Auli M. Unsupervised Cross-Lingual Representation Learning for Speech Recognition // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2021). 2021. pp. 2426–2430. DOI: 10.21437/Interspeech.2021-329.
8. Babu A., Wang C., Tjandra A., Lakhotia K., Xu Q., Goyal N., Singh K., Platen von P., Saraf Y., Pino J., Baevski A., Conneau A., Auli M. XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2022). 2022. pp. 2278–2282.
9. Pratap V., Tjandra A., Shi B., Tomasello P., Babu A., Kundu S., Elkahky A., Ni Zh., Vyas A., Fazel-Zarandi M., Baevski A., Adi Y., Zhang X., Hsu W.-N., Conneau A., Auli M. Scaling Speech Technology to 1,000+ Languages // Journal of Machine Learning Research. 2024. vol. 25. pp. 1–52.
10. Poth C., Sterz H., Paul I., Purkayastha S., Engländer L., Imhof T., Vulić I., Ruder S., Gurevych I., Pfeiffer J. Adapters: A unified Library for Parameter-Efficient and Modular Transfer Learning // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP'2023). 2023. pp. 149–160. DOI: 10.18653/v1/2023.emnlp-demo.13.
11. Chung Y.A., Zhang Y., Han W., Chiu C.-C., Qin J., Pang R., Wu Y. W2v-bert: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training // Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2021). 2021. pp. 244–250. DOI: 10.1109/ASRU51503.2021.9688253.
12. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'2019). 2019. vol. 1. pp. 4171–4186.
13. Chen S., Wang C., Chen Z., Wu Y., Liu S., Chen Z., Li J., Kanda N., Yoshioka T., Xiao X., Wu J., Zhou L., Ren S., Qian Y., Qian Y., Wu J., Zeng M., Yu X., Wei F. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing // IEEE Journal of Selected Topics in Signal Processing. 2022. vol. 16. no. 6. pp. 1505–1518.
14. Barrault L., Chung Y.A., Meglioli M.C., Dale D., Dong N., Duppenthaler M. et al. Seamless: Multilingual Expressive and Streaming Speech Translation // arXiv preprint arXiv:2312.05187. 2023.
15. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision // Proceedings of the 40th International Conference on Machine Learning 2022 (ICML'23). 2023. pp. 28492–28518.
16. Williams A., Demarco A., Borg C. The applicability of Wav2Vec 2.0 and Whisper for Low-Resource Maltese ASR // Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL'2023). 2023. pp. 39–43.
17. Kozhirbayev Z. Kazakh Speech Recognition: Wav2vec2.0 vs. Whisper // Journal of Advances in Information Technology. 2023. vol. 14. no. 6. pp. 1382–1389. DOI: 10.12720/jait.14.6.1382-1389.
18. Richardson F., Reynolds D., Dehak N. Deep Neural Network Approaches to Speaker and Language Recognition // IEEE Signal Processing Letters. 2015. vol. 22. no. 10. pp. 1671–1675. DOI: 10.1109/LSP.2015.2420092.

19. Winata G.I., Madotto A., Wu C.S., Fung P. Code-Switching Language Modeling using Syntax-Aware Multi-Task Learning // Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching (CALCS'2018). 2018. pp. 62–67. DOI: 10.18653/v1/W18-3207.
20. Adel H., Vu N.T., Kraus F., Schlippe T., Li H., Schultz T. Recurrent Neural Network Language Modeling for Code Switching Conversational Speech // Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2013). 2013. pp. 8411–8415. DOI: 10.1109/ICASSP.2013.6639306.
21. Ramanarayanan V., Pugh R., Suenderman-Oeft D. Automatic Turn-Level Language Identification for Code-Switched Spanish-English Dialog // Proceedings of 9th International Workshop on Spoken Dialogue System Technology (IWSDS'2019). 2019. vol. 579. pp. 51–61. DOI: 10.1007/978-981-13-9443-0_5.
22. Mustafa M.B., Yusoof M.A., Khalaf H.K., Abushariah A.A.R.M., Kiah M.L.M., Ting H.N., Muthaiyah S. Code-Switching in Automatic Speech Recognition: The Issues and Future Directions // Applied Sciences. 2022. vol. 12. no. 19. DOI: 10.3390/app12199541.
23. Çetinoğlu Ö., Schulz S., Vu N.T. Challenges of Computational Processing of Code-Switching // Proceedings of the Second Workshop on Computational Approaches to Linguistic Code Switching (CALCS'2016). 2016. pp. 1–11. DOI: 10.18653/v1/W16-5801.
24. Winata G., Aji A.F., Yong Z.X., Solorio T. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges // Findings of the Association for Computational Linguistics (ACL'2023). 2023. pp. 2936–2978. DOI: 10.18653/v1/2023.findings-acl.185.
25. Hsieh I.T., Wu C.H., Wang C.H. Acoustic and Textual Data Augmentation for Code-Switching Speech Recognition in Under-Resourced Language // IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC'2020). 2020. pp. 302–307.
26. Chang C.-T., Chuang S.-P., Lee H.-Y. Code-Switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2019). 2019. pp. 554–558. DOI: 10.21437/Interspeech.2019-3214.
27. Chan J.Y.C., Cao H., Ching P.C., Lee T. Automatic recognition of Cantonese-English Code-Mixing Speech // International Journal of Computational Linguistics and Chinese Language Processing. 2009. vol. 14. no. 3. pp. 281–304.
28. Adel H., Vu N.T., Kirchhoff K., Telaar D., Schultz T. Syntactic and Semantic Features for Code-Switching Factored Language Models // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015. vol. 23. no. 3. pp. 431–440. DOI: 10.1109/TASLP.2015.2389622.
29. Hermann K.M., Blunsom P. Multilingual Models for Compositional Distributed Semantics // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014. pp. 58–68. DOI: 10.3115/v1/P14-1006.
30. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave É., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised Cross-Lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'2020). pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.
31. Vüren van J., Niesler T. Improving N-best Rescoring in Under-Resourced Code-Switched Speech Recognition using Pretraining and Data Augmentation // Languages. 2022. vol. 7. no. 3. DOI: 10.3390/languages7030236.

32. Hono Y., Mitsuda K., Zhao T., Mitsui K., Wakatsuki T., Sawada K. Integrating Pre-Trained Speech and Language Models for End-to-End Speech Recognition // Findings of the Association for Computational Linguistics: ACL 2024. 2024. pp. 13289–13305. DOI: 10.18653/v1/2024.findings-acl.787.
33. Ogunremi T., Manning C.D., Jurafsky D. Multilingual Self-Supervised Speech Representations Improve the Speech Recognition of Low-Resource African Languages with Code Switching // arXiv preprint arXiv:2311.15077. 2023.
34. Hori T., Cho J., Watanabe S. End-to-end speech recognition with word-based RNN language models // Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT-2018). 2018. pp. 389–396. DOI: 10.1109/SLT.2018.8639693.
35. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček O., Qian Y., Schwarz P., Silovský J., Stemmer G., Veselý K. The Kaldi Speech Recognition Toolkit // Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2011). 2011. pp. 1–4.
36. Кипяткова И.С., Родионова А.П., Кагиров И.А., Крижановский А.А. Подготовка речевых и текстовых данных для создания системы автоматического распознавания карельской речи // Учёные записки Петрозаводского государственного университета. 2023. Т. 45. № 5. С. 89–98.
37. Povey D., Cheng G., Wang Y., Li K., Xu H., Yarmohammadi M., Khudanpur S. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks // Proceedings of The Annual Conference of the International Speech Communication Association (Interspeech'2018). 2018. pp. 3743–3747. DOI: 10.21437/Interspeech.2018-1417.
38. Saon G., Soltau H., Nahamoo D., Picheny M. Speaker Adaptation of Neural Network Acoustic Models using i-Vectors // Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2013). 2013. pp. 55–59. DOI: 10.1109/ASRU.2013.6707705.
39. Stolcke A., Zheng J., Wang W., Abrash V. SRILM at Sixteen: Update and Outlook // Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2011). 2011. pp. 5–9.
40. Enarvi S., Kurimo M. TheanoLM – An Extensible Toolkit for Neural Network Language Modeling // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2016). 2016. pp. 3052–3056. DOI: 10.21437/Interspeech.2016-618.
41. Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A. Large Vocabulary Russian Speech Recognition using Syntactico-Statistical Language Modeling // Speech Communication. 2014. vol. 56. pp. 213–228. DOI: 10.1016/j.specom.2013.07.004.
42. Wolf T., et al. Transformers: State-of-the-Art Natural Language Processing // arXiv preprint arXiv:1910.03771. 2019.
43. Grosz T., Getman Y., Al-Ghezi R., Rouhe A., Kurimo M. Investigating wav2vec2 Context Representations and the Effects of Fine-Tuning, a Case-Study of a Finnish Model // Proceedings of The Annual Conference of the International Speech Communication Association (Interspeech'2023). 2023. pp. 196–200. DOI: 10.21437/Interspeech.2023-837.
44. Hiovain-Asikainen K., Rosa de la J. Developing TTS and ASR for Lule and North Sámi Languages // Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL'2023). 2023. pp. 48–52. DOI: 10.21437/SIGUL.2023-11.
45. Paonessa C., Timmel V., Vogel M., Perruchoud D. Whisper Fine-Tuning for Swiss German: A Data Perspective // Proceedings of the 9th edition of the Swiss Text Analytics Conference. 2024. pp. 192.

Кипяткова Ирина Сергеевна — канд. техн. наук, доцент, старший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: автоматическое распознавание речи, нейронные сети. Число научных публикаций — 100. kipyatkova@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Кагиров Ильдар Амирович — научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: корпусная лингвистика, малоресурсные языки. Число научных публикаций — 40. kagirov@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Долгушин Михаил Дмитриевич — младший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: автоматическое распознавание речи, нейронные сети, математическая лингвистика. Число научных публикаций — 10. dolgushin.m@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Поддержка исследований. Работа выполнена при финансовой поддержке фонда РНФ (проект № 24-21-00276, https://rscf.ru/project/24-21-00276/).

DOI 10.15622/ia.24.2.9

I. KIPYATKOVA, I. KAGIROV, M. DOLGUSHIN

USE OF PRE-TRAINED MULTILINGUAL MODELS FOR KARELIAN SPEECH RECOGNITION

Kipyatkova I., Kagirov I., Dolgushin M. **Use of Pre-Trained Multilingual Models for Karelian Speech Recognition.**

Abstract. This paper presents an experimental study aimed at solving the problem of training speech recognition models under conditions of limited available speech and text data. Current approaches to this issue are discussed in detail, particularly the use of pre-trained multilingual models and data augmentation techniques. As part of this study, multilingual models based on Wav2Vec and Whisper were adapted to the Livvi dialect of the Karelian language, and an investigation into the use of an external language model to enhance recognition accuracy was conducted. The paper also describes a specially collected and prepared speech database and a basic recognition system developed using the Kaldi toolkit. Quantitative test results are provided as well, demonstrating the effectiveness of the chosen methods. For instance, Transformer-based models, particularly Wav2Vec, outperformed the baseline models trained using Kaldi software tools. Fine-tuning the Wav2Vec models reduced the word error rate to 24.73% on the validation set and 25.25% on the test set, while a combination of the Wav2Vec-BERT 2.0-based model with an external language model further reduced errors to 17.12% and 17.72%, respectively. This paper is primarily aimed at specialists in the field of automatic speech recognition for low-resource and Balto-Finnic languages. Additionally, the results of this work can be practically applied in field research involving Karelian text transcription. Future work includes expanding the database to improve model adaptation and enhance performance in real-world scenarios.

Keywords: low-resource languages, Karelian, code-switching, pre-trained models, machine learning, speech corpus.

References

1. Kipyatkova I., Kagirov I. [Automatic speech recognition system for Karelian]. Informatsionno-upravliaiushchie sistemy – Information and Control Systems. 2023. vol. 3. pp. 16–25. (In Russ.).
2. Kipyatkova I., Kagirov I. Deep Models for Low-Resourced Speech Recognition: Livvi-Karelian Case. Mathematics. 2023. vol. 11. no. 18. DOI: 10.3390/math11183814.
3. Kipyatkova I., Kagirov I. Phone Durations Modeling for Livvi-Karelian ASR. Proceedings 25th International Conference Speech and Computer (SPECOM 2023). Springer LNCS. 2023. vol. 14339. pp. 87–99. DOI: 10.1007/978-3-031-48312-7_7.
4. Kipyatkova I., Kagirov I., Dolgushin M., Rodionova A. Towards a Livvi-Karelian End-to-End ASR System. Proceedings 26th International Conference on Speech and Computer (SPECOM 2024). 2024. vol. 15299. pp. 57–68. DOI: 10.1007/978-3-031-77961-9_4.
5. Kipyatkova I.S., Kagirov I.A. [Analytical review of methods for solving data scarcity issues regarding elaboration of automatic speech recognition systems for low-resource languages]. Informatika i avtomatizacija – Informatics and Automation. 2022. vol. 21.no . 4. pp. 678–709. DOI: 10.15622/ia.21.4.2. (In Russ.).
6. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems. 2020. vol. 33. pp. 12449–12460.

7. Conneau A., Baevski A., Collobert R., Mohamed A., Auli M. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2021). 2021. pp. 2426–2430. DOI: 10.21437/Interspeech.2021-329.
8. Babu A., Wang C., Tjandra A., Lakhotia K., Xu Q., Goyal N., Singh K., Platen von P., Saraf Y., Pino J., Baevski A., Conneau A., Auli M. XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2022). 2022. pp. 2278–2282.
9. Pratap V., Tjandra A., Shi B., Tomasello P., Babu A., Kundu S., Elkahky A., Ni Zh., Vyas A., Fazel-Zarandi M., Baevski A., Adi Y., Zhang X., Hsu W.-N., Conneau A., Auli M. Scaling Speech Technology to 1,000+ Languages. Journal of Machine Learning Research. 2024. vol. 25. pp. 1–52.
10. Poth C., Sterz H., Paul I., Purkayastha S., Engländer L., Imhof T., Vulić I., Ruder S., Gurevych I., Pfeiffer J. Adapters: A unified Library for Parameter-Efficient and Modular Transfer Learning. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP'2023). 2023. pp. 149–160. DOI: 10.18653/v1/2023.emnlp-demo.13.
11. Chung Y.A., Zhang Y., Han W., Chiu C.-C., Qin J., Pang R., Wu Y. W2v-bert: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2021). 2021. pp. 244–250. DOI: 10.1109/ASRU51503.2021.9688253.
12. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'2019). 2019. vol. 1. pp. 4171–4186.
13. Chen S., Wang C., Chen Z., Wu Y., Liu S., Chen Z., Li J., Kanda N., Yoshioka T., Xiao X., Wu J., Zhou L., Ren S., Qian Y., Qian Y., Wu J., Zeng M., Yu X., Wei F. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing. 2022. vol. 16. no. 6. pp. 1505–1518.
14. Barrault L., Chung Y.A., Meglioli M.C., Dale D., Dong N., Duppenthaler M. et al. Seamless: Multilingual Expressive and Streaming Speech Translation. arXiv preprint arXiv:2312.05187. 2023.
15. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of the 40th International Conference on Machine Learning 2022 (ICML'23). 2023. pp. 28492–28518.
16. Williams A., Demarco A., Borg C. The applicability of Wav2Vec 2.0 and Whisper for Low-Resource Maltese ASR. Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL'2023). 2023. pp. 39–43.
17. Kozhirbayev Z. Kazakh Speech Recognition: Wav2vec2.0 vs. Whisper. Journal of Advances in Information Technology. 2023. vol. 14. no. 6. pp. 1382–1389. DOI: 10.12720/jait.14.6.1382-1389.
18. Richardson F., Reynolds D., Dehak N. Deep Neural Network Approaches to Speaker and Language Recognition. IEEE Signal Processing Letters. 2015. vol. 22. no. 10. pp. 1671–1675. DOI: 10.1109/LSP.2015.2420092.
19. Winata G.I., Madotto A., Wu C.S., Fung P. Code-Switching Language Modeling using Syntax-Aware Multi-Task Learning. Proceedings of the Third Workshop on

Computational Approaches to Linguistic Code-Switching (CALCS'2018). 2018. pp. 62–67. DOI: 10.18653/v1/W18-3207.
20. Adel H., Vu N.T., Kraus F., Schlippe T., Li H., Schultz T. Recurrent Neural Network Language Modeling for Code Switching Conversational Speech. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2013). 2013. pp. 8411–8415. DOI: 10.1109/ICASSP.2013.6639306.
21. Ramanarayanan V., Pugh R., Suenderman-Oeft D. Automatic Turn-Level Language Identification for Code-Switched Spanish-English Dialog. Proceedings of 9th International Workshop on Spoken Dialogue System Technology (IWSDS'2019). 2019. vol. 579. pp. 51–61. DOI: 10.1007/978-981-13-9443-0_5.
22. Mustafa M.B., Yusoof M.A., Khalaf H.K., Abushariah A.A.R.M., Kiah M.L.M., Ting H.N., Muthaiyah S. Code-Switching in Automatic Speech Recognition: The Issues and Future Directions. Applied Sciences. 2022. vol. 12. no. 19. DOI: 10.3390/app12199541.
23. Çetinoğlu Ö., Schulz S., Vu N.T. Challenges of Computational Processing of Code-Switching. Proceedings of the Second Workshop on Computational Approaches to Linguistic Code Switching (CALCS'2016). 2016. pp. 1–11. DOI: 10.18653/v1/W16-5801.
24. Winata G., Aji A.F., Yong Z.X., Solorio T. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. Findings of the Association for Computational Linguistics (ACL'2023). 2023. pp. 2936–2978. DOI: 10.18653/v1/2023.findings-acl.185.
25. Hsieh I.T., Wu C.H., Wang C.H. Acoustic and Textual Data Augmentation for Code-Switching Speech Recognition in Under-Resourced Language. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC'2020). 2020. pp. 302–307.
26. Chang C.-T., Chuang S.-P., Lee H.-Y. Code-Switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2019). 2019. pp. 554–558. DOI: 10.21437/Interspeech.2019-3214.
27. Chan J.Y.C., Cao H., Ching P.C., Lee T. Automatic recognition of Cantonese-English Code-Mixing Speech. International Journal of Computational Linguistics and Chinese Language Processing. 2009. vol. 14. no. 3. pp. 281–304.
28. Adel H., Vu N.T., Kirchhoff K., Telaar D., Schultz T. Syntactic and Semantic Features for Code-Switching Factored Language Models. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015. vol. 23. no. 3. pp. 431–440. DOI: 10.1109/TASLP.2015.2389622.
29. Hermann K.M., Blunsom P. Multilingual Models for Compositional Distributed Semantics. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014. pp. 58–68. DOI: 10.3115/v1/P14-1006.
30. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave É., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised Cross-Lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'2020). pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.
31. Vüren van J., Niesler T. Improving N-best Rescoring in Under-Resourced Code-Switched Speech Recognition using Pretraining and Data Augmentation. Languages. 2022. vol. 7. no. 3. DOI: 10.3390/languages7030236.
32. Hono Y., Mitsuda K., Zhao T., Mitsui K., Wakatsuki T., Sawada K. Integrating Pre-Trained Speech and Language Models for End-to-End Speech Recognition. Findings

of the Association for Computational Linguistics: ACL 2024. 2024. pp. 13289–13305. DOI: 10.18653/v1/2024.findings-acl.787.

33. Ogunremi T., Manning C.D., Jurafsky D. Multilingual Self-Supervised Speech Representations Improve the Speech Recognition of Low-Resource African Languages with Code Switching. arXiv preprint arXiv:2311.15077. 2023.
34. Hori T., Cho J., Watanabe S. End-to-end speech recognition with word-based RNN language models. Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT-2018). 2018. pp. 389–396. DOI: 10.1109/SLT.2018.8639693.
35. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček O., Qian Y., Schwarz P., Silovský J., Stemmer G., Veselý K. The Kaldi Speech Recognition Toolkit. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2011). 2011. pp. 1–4.
36. Kipyatkova I.S., Rodionova A.P., Kagirov I.A., Krizhanovsky A.A. [Speech and text data preparation for developing of an automatic speech recognition system for the Karelian language]. Uchjonye zapiski Petrozavodskogo gosudarstvennogo universiteta – Proceedings of Petrozavodsk State University. 2023. vol. 45. no. 5. pp. 89–98. (In Russ.).
37. Povey D., Cheng G., Wang Y., Li K., Xu H., Yarmohammadi M., Khudanpur S. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. Proceedings of The Annual Conference of the International Speech Communication Association (Interspeech'2018). 2018. pp. 3743–3747. DOI: 10.21437/Interspeech.2018-1417.
38. Saon G., Soltau H., Nahamoo D., Picheny M. Speaker Adaptation of Neural Network Acoustic Models using i-Vectors. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2013). 2013. pp. 55–59. DOI: 10.1109/ASRU.2013.6707705.
39. Stolcke A., Zheng J., Wang W., Abrash V. SRILM at Sixteen: Update and Outlook. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'2011). 2011. pp. 5–9.
40. Enarvi S., Kurimo M. TheanoLM – An Extensible Toolkit for Neural Network Language Modeling. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'2016). 2016. pp. 3052–3056. DOI: 10.21437/Interspeech.2016-618.
41. Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A. Large Vocabulary Russian Speech Recognition using Syntactico-Statistical Language Modeling. Speech Communication. 2014. vol. 56. pp. 213–228. DOI: 10.1016/j.specom.2013.07.004.
42. Wolf T., et al. Transformers: State-of-the-Art Natural Language Processing. arXiv preprint arXiv:1910.03771. 2019.
43. Grosz T., Getman Y., Al-Ghezi R., Rouhe A., Kurimo M. Investigating wav2vec2 Context Representations and the Effects of Fine-Tuning, a Case-Study of a Finnish Model. Proceedings of The Annual Conference of the International Speech Communication Association (Interspeech'2023). 2023. pp. 196–200. DOI: 10.21437/Interspeech.2023-837.
44. Hiovain-Asikainen K., Rosa de la J. Developing TTS and ASR for Lule and North Sámi Languages. Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL'2023). 2023. pp. 48–52. DOI: 10.21437/SIGUL.2023-11.
45. Paonessa C., Timmel V., Vogel M., Perruchoud D. Whisper Fine-Tuning for Swiss German: A Data Perspective. Proceedings of the 9th edition of the Swiss Text Analytics Conference. 2024. pp. 192.

Kipyatkova Irina — Ph.D., Associate Professor, Senior researcher, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: automatic speech recognition, neural networks. The number of publications — 100. kipyatkova@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Kagirov Ildar — Research fellow, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: corpus linguistics, low-resource languages. The number of publications — 40. kagirov@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Dolgushin Mikhail — Junior research fellow, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: automatic speech recognition, neural networks, mathematical linguistics. The number of publications — 10. dolgushin.m@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

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