УДК 004.522 **DOI** 10.15622/ia.21.4.2

И.С. Кипяткова, И.А. Кагиров

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ РЕШЕНИЯ ПРОБЛЕМЫ МАЛЫХ НАБОРОВ ДАННЫХ ПРИ СОЗДАНИИ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ МАЛОРЕСУРСНЫХ ЯЗЫКОВ

Кипяткова И.С., Кагиров И.А. **Аналитический обзор методов решения проблемы малых наборов данных при создании систем автоматического распознавания речи для малоресурсных языков.**

Аннотация. В статье рассматриваются основные методы решения проблемы малых наборов обучающих данных для создания автоматических систем распознавания речи для так называемых малоресурсных языков. Рассматривается понятие малоресурсных языков и формулируется рабочая дефиниция на основании ряда работ по этой тематике. Определены основные трудности, связанные с применением классических схем автоматического распознавания речи к материалу малоресурсных языков, и очерчен круг основных методов, использующихся для решения обозначенных проблем. В статье подробно рассматриваются методы аугментации данных, переноса знаний и сбора речевого материала. В зависимости от конкретной задачи, выделяются методы аугментации аудиоматериала и текстовых данных, переноса знаний и мультизадачного обучения. Отдельный раздел статьи посвящен существующему информационному обеспечению, базам данных и основным принципам их организации с точки зрения работы с малоресурсными языками. Делаются выводы об оправданности методов аугментации данных и переноса знаний для языков с минимальным информационным обеспечением. В случае полного отсутствия данных для конкретного языка и родительских моделей структурно схожих языков предпочтительным вариантом является сбор новой базы данных, в том числе, при помощи краудсорсинга. Многозадачные модели переноса знаний оказываются эффективными в том случае, если исследователь располагает набольшими наборами данных. Если доступны данные по языку с достаточными ресурсами, предпочтительной является работа с языковой парой. Сделанные в результате данного обзора выводы в дальнейшем предполагается применить при работе с малоресурсным карельским языком, для которого авторы статьи создают систему автоматического распознавания речи.

Ключевые слова: малоресурсные языки, аугментация речевых данных, перенос знаний, машинное обучение, языковые корпуса.

1. Введение. Существующие в настоящее время системы автоматической обработки естественных языков (распознавания речи, анализа тональности, автоматического перевода) построены при помощи технологий машинного обучения, требующих большого количества данных для эффективной работы. Однако языков с достаточной степенью изученности и готовыми большими наборами данных сравнительно мало. По разным оценкам, в мире насчитывается от 5 до 7 тысяч языков, и языками с достаточными ресурсами могут быть признаны только около 20 из них [1, 2]. Большинство языков мира описаны и документированы недостаточно хорошо; более того,

данные по ним часто недоступны. Такие языки принято называть малоресурсными, и применение к ним современных методов обработки естественных языков зачастую невозможно. Осознание актуальности этой проблемы неоднократно находило отражение в научных публикациях [3, 4].

В настоящей статье содержится обзор основных методов, связанных с применением речевых технологий в контексте малоресурсных языков. Из-за специфики темы в фокусе данной статьи оказываются как информационное обеспечение малоресурсных языков, то есть наборы данных, так методика их создания. Целью настоящей статьи является систематизация основных методов получения и подготовки данных для применений в системах распознавания речи (и шире – автоматической обработки естественных языков) для малоресурсных языков. В статье выполнено выделение основных методов, применяемых для решения проблемы малых и/или недостаточных наборов данных для малоресурсных языков, представлена классификация методов по типам задач, для которых они применяются, проведен их сравнительный анализ. Кроме того, в статье подробно рассматривается понятие малоресурсного языка, которое является ключевым для определения набора методов, пригодных для работы с конкретным языковым материалом.

Статья имеет следующую структуру: после настоящего Введения дается краткое определение малоресурсных языков. Далее приводится общий обзор проектов и наборов данных, связанных с малоресурсными языками. Следующие разделы посвящены основным подходам к малоресурсным языкам, в том числе методам сбора данных, аугментации данных и переноса знаний при обучении моделей. Наконец, в Заключении делаются выводы о положении вещей в рассматриваемой области.

2. Малоресурсные языки: определение и проблемы автоматической обработки. В современной традиции обработки естественных языков под термином «малоресурсные языки» понимаются языки с малым объемом электронных ресурсов, доступных для обработки. Впервые термин «малоресурсные языки» (англ. low-resource languages, under-resourced languages) был предложен в работах [5, 6]. Одновременно был выдвинут ряд критериев, на основании которых тот или иной язык может быть отнесен к малоресурсным: наличие письменности, доступность языковых материалов в сети Интернет, существование лингвистических описаний языка, электронных (двуязычных) словарей и параллельных корпусов и т.п. В этих же работах были предложены системы оценки

доступности ресурсов для конкретного языка, основанные на вычислении среднего значения коэффициентов, присваиваемых каждому ресурсу.

В более поздних работах [7, 8] содержание термина «малоресурсные языки» было еще расширено: учитываются такие факторы, как низкий социальный статус языка и его малоизученность. Тем не менее, на практике [9] основным критерием для отнесения какого-либо языка к малоресурсным является низкий для решения конкретной задачи объем электронных языковых данных, которым располагают исследователи.

Обыкновенно существует прямая взаимосвязь между собственно уровнем развития информационных технологий для языка и степенью лингвистической изученности языкового материала. Кроме того, чаще всего малоресурсными оказываются миноритарные и вымирающие языки [10]. Следует подчеркнуть, что описанная ситуация является типичной, но не обязательной; так, ряд хорошо изученных языков, имеющих официальный статус, может быть отнесен к малоресурсным (казахский или белорусский), в то время как некоторые миноритарные языки таковыми не являются (баскский язык).

Малоресурсные языки, как и любые естественные языки, представляют большой интерес не только для лингвистов. В статье [8] отмечается, что только в Африке и в Индии существует около 2 тысяч малоресурсных языков, на которых разговаривают порядка 2,5 миллиарда человек. Создание инструментов для естественной коммуникации с носителями этих языков может помочь в решении ряда проблем экономического, культурного и экологического характера.

Низкий уровень развития речевых технологий для малоресурсных языков объясняется рядом причин. В первую очередь, задача документации и формализации структур малоизученного языка сама по себе достаточно трудоемка с научной точки зрения. Кроме того, она сопряжена с высокими финансовыми затратами на подготовку необходимых инструментов (речевых корпусов, программного обеспечения). Также важно то, что существующие на сегодня способы моделирования и создания комплексов распознавания речи направлены на решение узкого круга задач и не учитывают особенности работы с малоресурсными языками.

Проблемам малоресурсных языков, созданию наборов данных и адаптации моделей для автоматической обработки малоресурсных языков посвящен ряд конференций и семинаров. В первую очередь,

это International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU), Collaboration and Computing for Under-Resourced Languages (CCURL), Language Resources (LRs) and Evaluation for Language Technologies (LREC), также тематика малоресурсных языков часто фигурирует в повестке других мероприятий, посвященных автоматической обработке естественных языков (Association for Computational Linguistics (ACL) European Chapter of the Association for Computational Linguistics (EACL) и т.п.).

В России существует свыше 150 языков[*], многие из которых являются малоресурсными и не имеют собственных корпусов. Тем не менее, за последние годы появляются работы, в которых предлагаются методы и системы для автоматической обработки этих языков, предпринимаются попытки создания корпусов. В качестве примера можно привести работы [11] (татарский язык), [12] (чеченский язык), [13] (вепсский и карельский языки).

3. Проблемы автоматической обработки малоресурсных языков и основные методы их решения. Для удобства читателя на рисунке 1 представлена общая архитектура системы автоматического распознавания слитной речи. Применение подобных архитектур к материалу малоресурсных языков обнаруживает ряд проблем на различных этапах распознавания, отраженных на схеме.

Как правило, к основным компонентам подобных систем распознавания речи относятся акустическая, лексическая и языковая модели. Для обучения акустических моделей необходим речевой корпус, для обучения модели языка – текстовый. Кроме того, необходимо создать транскрипции для слов, которые будут использоваться в системе. Стоит отметить, что в настоящее время развивается интегральный (англ. end-to-end) подход к построению систем распознавания речи, который заключается в том, что только одна модель генерирует необходимые выходные данные без использования других компонентов [14].

Как было отмечено в самом начале этой статьи, традиционные подходы к автоматической обработке естественного языка требуют больших объемов обучающих данных, однако для малоресурсных языков такой объем данных недоступен. Недоступность данных и знаний по языку – это комплексная проблема, она актуальна для всех уровней обработки языка: собственно фонетического, лексического и грамматического [15]. Особенно остро проблема недостаточного объема обучающих данных стала проявляться с развитием

[*] Ethnologue: Languages of the World. Russian Federation. URL: https://www.ethnologue.com/country/RU/languages (дата обращения: 22.06.2022.)

искусственных нейронных сетей (ИНС), которые позволяют повысить точность автоматического распознавания речи по сравнению с другими подходами (в частности, скрытыми марковскими моделями для акустического моделирования, и *n*-граммами – для языкового), однако требуют больший объем обучающих данных, поэтому в данной статье главным образом приведены методы, применяемые для обучения нейросетевых моделей. Однако также описываются методы аугментации (увеличения объема) данных, которые могут использоваться и для обучения систем, построенных не на нейронных сетях.

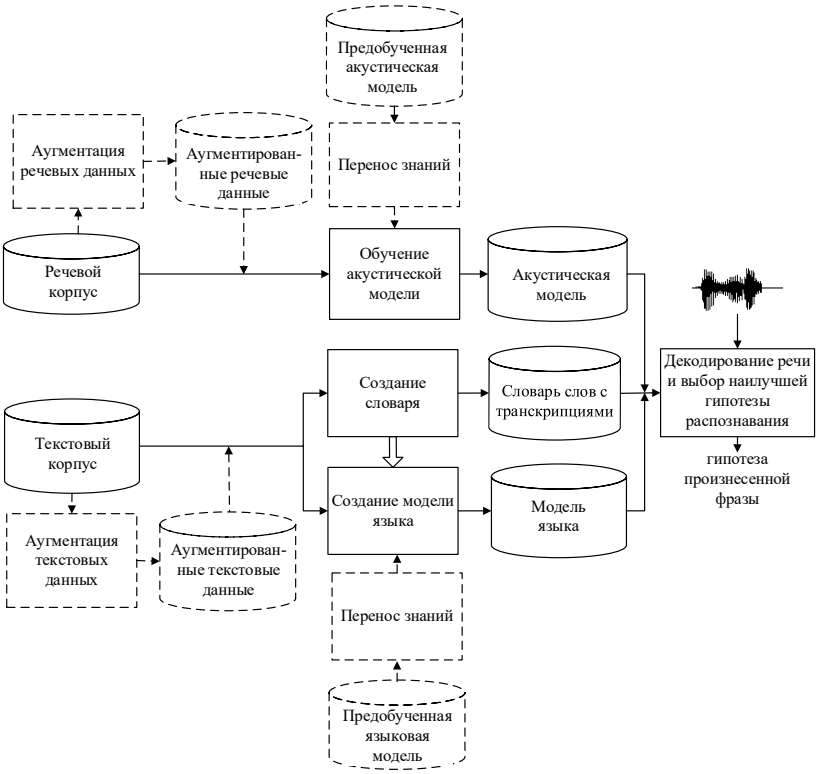

Рис. 1. Общая архитектура системы автоматического распознавания слитной речи

На рисунке пунктирной линией отмечены модули, которые могут входить в состав системы распознавания речи для малоресурсных языков:

1) аугментация данных (аудиоданные, текстовые данные);
2) перенос знаний (в том числе многозадачное обучение).

Строго говоря, еще одним эффективным инструментом получения обучающих данных является сбор собственного корпуса, однако этот метод требует больших временных, человеческих и зачастую финансовых затрат.

Также следует отметить, что при работе с малоресурсными языками исследователю приходится решать ряд других, специфических для этой области проблем (нестандартная письменность или ее отсутствие, типологическая уникальность фонетических кластеров, создание словаря транскрипций и т.п.), однако данный обзор ограничивается только проблемой малого объема данных, поскольку она является основной для всех малоресурсных языков. Последующие разделы будут посвящены методам решения этой проблемы.

4. Методы сбора и транскрибирования речевых данных. Самым надежным способом получения языковых данных для любого естественного языка является запись собственного корпуса. Сбор таких корпусов представляет собой задачу, сопряженную с рядом трудностей (географическая удаленность, доступность дикторов и текстов, социальные проблемы). За последние годы был собран достаточно большой массив данных по малоресурсным языкам, однако общее количество таких корпусов оценивается в несколько десятков [16].

Одним из способов сбора и разметки данных для обучения моделей является краудсорсинг, при этом можно выделить три типа краудсорсинга [17]: 1) наемный труд, к котором привлекаемые люди получают оплату за свою работу; 2) игры с целью (англ. games with a purpose; GWAP), где задача представлена как игра; 3) привлечение волонтеров.

В работе [18] авторы описывают созданное ими приложение для сбора речевых данных путем краудсорсинга. Приложение имеет клиент-серверную архитектуру. Клиентское приложение ставится на смартфоны под управлением операционной системы Андроид. Клиентское приложение связывается с сервером, пользователь читает предложения, которые показываются ему на экране, по окончании записи речевые данные отравляются на сервер.

Среди крупнейших проектов по сбору данных малоресурсных языков стоит назвать такие, как GlobalPhone, LORELEI, REFLEX-LCTL и IARPA BABEL.

Корпус GlobalPhone [19] – это многоязычный (20 языков) корпус данных, разработанный в сотрудничестве с Технологическим институтом Карлсруэ (KIT). Полный корпус данных включает: 1) аудио/речевые данные, т. е. высококачественные записи устных высказываний, прочитанных носителями языка; 2) соответствующие транскрипции; 3) словари произношения, охватывающие словарный запас расшифровок; 4) базовые n-граммные языковые модели. Первые два называются базой данных речи и текста GlobalPhone (GP-ST), третий – словарями GlobalPhone (GP-Dict), а последний – языковыми моделями GlobalPhone (GP-LM). Весь корпус GlobalPhone представляет собой многоязычную базу данных расшифрованной речи на уровне слов для разработки и оценки систем обработки речи с большим словарем на самых распространенных языках мира. Корпус GlobalPhone создан единообразно для всех включенных в нее языков в плане объема текста и речи для каждого языка (100 дикторов на язык), качества записей и аннотаций.

В проекте LORELEI (Low Resource Languages for Emergent Incidents) [20], созданном Консорциумом языковых данных, представлены 35 языков; языки делятся на репрезентативные (23) и прочие (12). Репрезентативные языки используются для обеспечения широкого типологического охвата, а оставшиеся 12 языков были выбраны для обеспечения разработки и тестирования возможностей системы. В LORELEI языки организованы в так называемые «репрезентативные языковые пакеты», интегрированные в общую систему аннотаций. Репрезентативный языковой пакет для LORELEI содержит одноязычный текст, параллельный текст, несколько типов аннотаций, инструменты для обработки текста, сегментации и маркировки языковых единиц, а также словари и грамматические правила. Языковые пакеты LORELEI были собраны специально в рамках предметной темы «ситуационная осведомленность в непредвиденных ситуациях», что обеспечивает проекту определенную социальную значимость.

Проект REFLEX-LCTL (Research on English and Foreign Language Exploitation-Less Commonly Taught Languages) [21], спонсируемый правительством Соединенных Штатов, представлял собой другую попытку создания базовых языковых ресурсов для нескольких малоресурсных языков. В рамках этого проекта были созданы языковые пакеты для 19 языков. Этот подход сходен с

подходом, предложенным в LORELEI, однако, в отличие от LORELEI, пакеты языковых данных не объединены в рамках общей системы аннотаций.

Одним из крупнейших проектов по сбору языковых данных является IARPA BABEL [†] – проект, в котором приняли участие исследователи из различных стран Европы. Результатом проекта BABEL является многоязычная база данных, включающая транскрибированные записи речи на нескольких десятках малоресурсных разноструктурных языков из различных географических ареалов. Как утверждают сами создатели базы данных, основной целью проекта является разработка технологий распознавания речи для широкого спектра языков, в частности, методов моделирования малоресурсных языков в условиях малых наборов обучающих данных и их низкого качества (шумы, несбалансированность данных).

Следующим этапом после записи речевых данных является их транскрибирование, поскольку для обучения акустических моделей необходимо иметь расшифровку речевых записей в текстовом виде. Для малоресурсных языков транскрибирование речевых данных может выполняться путем краудсорсинга, при котором для разметки данных привлекают людей, не говорящих на языке, речевые данные которого необходимо транскрибировать, при этом люди записывают то, что они слышат, как звуки собственного языка. Получаемая таким образом транскрипция, называемая «несоответствующей» (англ. "mismatched"), преобразуется затем в транскрипцию для целевого языка, называемою вероятностной транскрипцией. Такой подход к транскрибированию был исследован в работе [22] для вьетнамского языка, при этом «несоответствующие» транскрипции были сделаны носителями путунхуа. Также в работе было исследовано совместное использование «соответствующих» и «несоответствующих» транскрипций и было показано, что использование «несоответсвующих» транскрипций может повысить точность распознавания при недостатке «соответствующих» транскрипций.

Также применение вероятностных транскрипций исследовалось и в работе [23]. Транскрипции создавались следующим образом: 1) самообучение (система распознавания речи, предобученая на других языках, использовалась для транскрибирования данных целевого языка); 2) "mismatched" краудсорсинг; 3) использование электроэнцефалограммы (ЭЭГ): суть метода состоит в том, что

[†] BABEL // Intelligence Advanced Research Projects Activity. URL: https://www.iarpa.gov/research-programs/babel (дата обращения: 22.06.2022.)

вначале, пока человек слушает речь на родном ему языке, записывается его ЭЭГ, выполняется обучение классификаторов признаков на размеченных по типам (например, фрикативный, сонорный) фонах. Затем человек слушает речь на целевом языке, и результаты его ЭЭГ используются для оценки распределения вероятности по набору фонов целевого языка.

5. Аугментация обучающих данных. Аугментация данных – это набор методов, которые используются для создания дополнительных данных либо путем модификации существующих данных, либо путем добавления данных из сторонних источников для использования при обучении модели.

5.1. Аугментация речевых данных. Аугментация речевых данных может проводиться путем изменения высоты голоса, темпа речи, громкости речи, наложения шума, модификации признаков, извлеченных из речевого сигнала, а также синтеза речи [24]. Аугментация данных путем изменения речевого сигнала описана в работе [25], где было предложено выполнить изменение темпа речи путем умножения исходной скорости на коэффициенты 0,9, 1,0 и 1,1, что позволило снизить количество неправильно распознанных слов (англ. word error rate; WER) на 2% для путунхуа. Аугментация путем добавления к речевым признакам случайных значений была выполнена в работе [26]. Кроме того, может быть выполнена аугментация не самого речевого сигнала, а его спектрограммы, подобный подход был применен в работе [27] для интегрального распознавания речи.

В некоторых работах применяется сразу несколько типов аугментации. В частности, в работе [28] предложена двухступенчатая аугментация речевых данных. На первом этапе для повышения робастности акустических моделей к исходным речевым данным был добавлен случайный шум, а также выполнено изменение темпа речи. На втором этапе происходит аугментация признаков, полученных путем адаптации к голосу диктора.

Также для аугментации может использоваться технология преобразования голоса (англ. voice conversion), которая состоит в модификации исходной аудиозаписи голоса диктора в голос другого диктора (целевой голос) без изменения лексического содержания речи [29]. Чаще всего для этого используют генеративно-состязательные сети (англ. Generative Adversarial Network (GAN)) [30], а также их модификации, например, Wasserstein GAN, StarGAN. В работе [31] для непараллельного преобразования голоса предложен метод VAW-GAN, объединяющий условный вариационный автоэнкодер (англ. conditional

variational autoencoder; C-VAE) для моделирования акустических признаков речи от каждого диктора и Wasserstein GAN для синтезирования голоса другого диктора. Архитектура StarGAN используется в работе [32], в которой описан метод преобразования речи, названный StarGAN-VC.

Несколько типов аугментации было применено в работе [33] для распознавания турецкой речи, в который авторы исследовали такие виды аугментации, как изменения скорости речи, громкости, совместное изменение скорости и громкости, синтез речи (исследовалась система преобразования речи в текст от Google и интегральная система синтеза турецкой речи на основе сверточных глубоких сетей). Кроме того, были применены различные комбинации из описанных выше методов. Наилучший результат был получен при совместном применении всех методов, при этом снижение WER составило 14,8%.

Еще одним методом аугментации речевых данных является синтез речи. В большинстве современных работ для синтеза речи используется нейросетевая модель Tacatron 2 [34] от компании Google. В частности, такой метод аугментации был выполнен в работе [35] для синтезирования детской речи на языке панджаби. Кроме того, в данной работе была выполнена аугментация за счет модификации формант в речевых записях корпуса взрослой речи. В работе [36] синтез речи использовался для аугментации речевых данных для разработки интегральной системы распознавания речи, что позволило существенно снизить WER, кроме того, был применен SpecAugment, что позволило получить дополнительное снижение WER. Недостатком такого метода аугментации является необходимость наличия речевых данных для обучения синтезатора речи, при их недостатке качество синтезируемой речи может быть неудовлетворительным. Так, авторы работы [37] не смогли получить увеличения точности распознавания при добавлении синтезированных данных к реальным при обучении акустических моделей, при этом авторы использовали статистический параметрический синтез речи. Также авторы выполнили синтез речи с помощью моделей Tacatron 2 и WGANSing (синтезатора, основанного на генеративных состязательных нейронных сетях), однако качество получившейся речи было настолько плохим, что авторы не стали проводить эксперименты с использованием этих данных. Плохое качество синтеза авторы объясняют недостатком данных для обучения.

5.2. Аугментация текстовых данных. Методы аугментации текстовых данных в контексте обработки естественных языков можно

разделить на две группы: с применением данных из других языков и без иноязычных данных.

5.2.1. Аугментация на материале одного языка. При сборе текстового материала для создания модели языка может возникнуть ситуация, когда текстового материала соответствующей предметной области мало, но есть текстовые данные для других предметных областей рассматриваемого языка. Особенно такая ситуация актуальна для малоресурсных языков. В этом случае можно обучить несколько моделей языка на текстах по каждой из имеющихся предметных областей, а затем выполнить их линейную интерполяцию, при этом коэффициент интерполяции подбирается так, чтобы получить минимальный коэффициент неопределенности (англ. perplexity) на отладочном текстовом корпусе [38].

В работе [39] выделены основные способы создания нейросетевой модели языка для заданной предметной области: 1) выбор данных из предметной области с большей вероятностью в ходе обучения (например, каждую эпоху использовать все данные предметной области и случайный поднабор остальных данных); 2) сортировка обучающих данных таким образом, чтобы данные непредметной области подавались на вход в начале обучение, а данные предметной области – в конце обучения, чтобы они имели большее влияние на модель; 3) обучение модели на данных непредметной области с последующей адаптацией модели к данным предметной области, например, посредством добавления к нейронной сети еще одного слоя и обучения его на данных предметной области, с предварительной фиксацией параметров остальных слоев; 4) большая часть параметров модели является общей для всех предметных областей, а некоторая часть параметров резервируется для конкретных предметных областей; таким образом, создается одна модель для различных предметных областей.

Распространенным методом текстовой аугментации является аугментация на основе замены слов или фраз, при которой данные расширяются за счет замены одних слов или словосочетаний синонимичными единицами. Типологически сходными методами аугментации является применение сокращений, использование векторных и контекстных представлений слов, однако эффективность их применения к материалу малоресурсных языков сильно ограничена наличием описаний и/или компьютерных моделей для конкретных языков [40].

В работе [41] исследуется аугментация текстовых данных на трех уровнях: символы (вставка/удаление/замена символа и

перестановка символов), слова (вставка/удаление/ перестановка слов, замена слова на синоним) и синтаксис. Самым сложным является синтаксический уровень, для работы с которым необходима разметка корпуса.

В работе [42] предложен метод аугментации размеченных текстовых данных, названный контекстной аугментацией, который состоит в том, что предварительно обученная языковая модель применяется для генерации замен слов в предложении на основе их контекста. В [42] в качестве такой модели использовалась двунаправленный рекуррентная ИНС. В работе [43] для контекстной аугментации данных использовались предварительно обученные модели на базе архитектуры трансформер: GPT-2, BERT и BART. Также аугментация путем замены слов была выполнена в работе [44] для адаптации модели к конкретной предметной области. В предложенном методе вначале выполняется обучение модели замены слов на основе двунаправленной сети с долгой кратковременной памятью (англ. Long Short-Term Memory; LSTM), аналогичной предложенной в работе [42], на размеченных исходных данных и данных заданной предметной области. Затем с использованием созданной модели выполняется преобразование предложений из текстового корпуса исходной предметной области в сеть спутывания (англ. confusion network), которая включает в себя возможные варианты последовательности слов для целевой предметной области. Затем с помощью модели языка на основе LSTM из сети спутывания выбираются предложения, которые, как подразумевается, являются грамматически корректными.

5.2.2. Аугментация с привлечением иноязычных данных. В том случае, если для конкретного малоресурсного языка существуют параллельные корпуса и/или системы машинного перевода, возможно применение методов аугментации с привлечением данных из другого языка. Простейшим методом является обратный перевод – процесс перевода одноязычного корпуса на целевой язык с помощью уже существующей системы машинного перевода с последующим обратным переводом на исходный язык. Затем полученные предложения исходного языка вместе с соответствующими предложениями целевого языка используются для построения искусственного (т.е. аугментированного) параллельного корпуса [45]. В некоторых случаях перевод с целевого языка на исходный осуществляется несколько раз (итеративный обратный перевод). Данные, полученные с использованием методики обратного перевода, обыкновенно характеризуются большим количеством ошибок, чем

исходные параллельные данные, особенно если система машинного перевода, используемая для создания синтетических данных, несовершенна.

Основная трудность, сопряженная с этой методикой, состоит в том, что для ее осуществления необходимо наличие системы машинного перевода для данной языковой пары. В ряде работ отмечается, что эффективность обратного перевода зависит от многих факторов, таких как соотношение параллельных данных, а также структурное соответствие параллельных и одноязычных данных [48, 49]. Попытки использовать языковые данные языков с достаточными ресурсами также сильно зависят от родства внутри языковой пары [50] или наличия двуязычных словарей [51].

В том случае, когда два параллельных корпуса не представляют собой результат прямого перевода, но все же принадлежат к одной предметной области (хорошим примером являются статьи из Википедии на разных языках), их также можно использовать для аугментации данных. Основным инструментом в данном случае является векторное представление предложений и словосочетаний в текстах. Тем не менее, для малоресурсных языков этот метод не всегда применим либо в силу отсутствия собственно параллельных корпусов, либо из-за отсутствия адекватных языковых моделей для конкретных языковых пар.

В таблице 1 показано относительное снижение ошибки распознавание слов (WER), полученное различными исследователями за счет применение методов аугментации данных при обучении систем распознавания речи.

Наиболее эффективными методами аугментации аудиоданных могут быть признаны синтез речи и изменение спектрограммы. Несмотря на относительную простоту и низкие вычислительные затраты, аугментация при помощи изменения темпа речи и высоты голоса, а также добавление случайных значений к речевым признакам не обеспечивают значительного улучшения точности распознавания слов при применении в системах распознавания речи на малоресурсных языках.

Таблица 1. Результаты применения методов аугментации обучающих данных для обучения систем автоматического распознавания речи

Методы аугментации	Изменяемые параметры	Работы	Относительное сокращение ошибки распознавания слов (WER, %)
Речевые данные	высота голоса/ темп речи/ громкость речи	[25]	2,00
	добавление случайных значений к речевым признакам	[26]	0,30
	изменение спектрограммы	[27]	27-46 (в зависимости от используемых данных и применения модели языка)
	преобразование голоса (voice conversion)	[46]	31,36
	синтез речи	[47]	60,40
Текстовые данные	использование текстовых данных другой предметной области	[39]	2,70 – 3,90 (в зависимости от объема данных)

6. Метод переноса знаний. Метод переноса знаний (англ. transfer learning) — это метод обучения ИНС, когда знания нейросети, которая была обучена на одной задаче, переносятся на другую задачу. Этот метод используется в том случае, если обучающих данных для целевой задачи мало, но имеется большой обучающий корпус для другой смежной задачи.

Существует несколько подходов к выполнению переноса знаний, самым простым является адаптация существующей модели к новым данным [52], для чего могут использоваться методы максимального правдоподобия для нахождения параметров линейной регрессионной модели (англ. maximum likelihood linear regression (MLLR)) и адаптация с использование апостериорного максимума (англ. maximum a posteriori (MAP)) [53]. Метод MLLR изначально был предложен для адаптации к голосу диктора, он позволяет выполнять адаптацию акустических моделей без повторного обучения параметров модели. MAP используется для адаптации к голосу диктора дикторонезависимых моделей по небольшому набору адаптационных данных. Так, например, в работе [54] методы MLLR и MAP

использовались для адаптации акустической модели, обученной на данных фламандских диалектов нидерландского языка, к близкородственному языку африкаанс. Также в данной работе исследовалось применение гетероскедастического линейного дискриминантного анализа (англ. heteroscedastic linear discriminant analysis (HLDA)) и адаптация к речи диктора.

Особую эффективность метод переноса знаний показал при обучении глубоких ИНС. С точки зрения систем распознавания речи, идея переноса знаний основана на том, что признаки, выделяемые при обучении нижними слоями ИНС, не зависят от языка, при этом специфичные для языка признаки выделяются верхними слоями ИНС. Обучение модели для малоресурсного языка с использованием метода переноса знаний выполняется в два этапа. Первым этапом является обучение родительской модели языка (англ. parent model) на большом наборе данных (нецелевых данных), которая впоследствии используется для инициализации весов в дочерней модели, обученной на материале малоресурсного языка (целевых данных). Очевидным преимуществом такого подхода является снятие проблемы малых данных для малоресурсного целевого языка. Родительская модель может быть обучена как на одном языке, так и на нескольких языках. Общая схема применения метода переноса знаний для обучения модели для малоресурсных языков показана на рисунке 2.

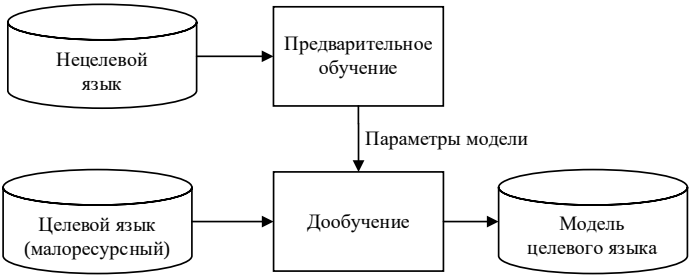

Рис. 2. Схема применения метода переноса знаний для обучения модели для малоресурсных языков

Метод переноса знаний был применен в работе [55] при обучении акустической модели для амхарского языка. В качестве нецелевых данных для предварительного обучения модели использовались речевые данные на английском и путунхуа. Применение метода переноса знаний позволило снизить WER с 38,72% до 24,50%.

В работе [56] применен метод переноса знаний для обучения системы распознавания речи для малоресурсного языка североамериканских индейцев сенека. Размер корпуса составлял всего 10 часов. К корпусу были применены три метода аугментации данных: 1) изменение скорости речи и частоты основного тона; 2) преобразование голоса с помощью метода StarGAN-VC; 3) преобразование голоса с помощью VAWGAN. Обучение нейросетевых акустических моделей происходило за 3 этапа. На первом этапе было выполнено предобучение модели на 960 часах английской речи из корпуса LibriSpeech. На втором этапе веса модели были проинициализированы значениями, полученными в ходе первого этапа обучения модели, и было произведено обучение модели на корпусе, включающем в себя 10 часов речи и аугментированные данные. На третьем этапе веса нейронной сети были проинициализированы значениями, полученными на втором этапе, и было выполнено обучение модели на исходных неаугментированных данных, чтобы исключить влияние артефактов аугментации.

В работе [57] исследуется влияние заморозки параметров нижних слоев ИНС при выполнении переноса знаний при обучении интегральной системы автоматического распознавания речи. Авторы обнаружили, что заморозка параметров нижних слоев позволяет повысить точность распознавания и скорость обучения. Существенное увеличение точности было получено при заморозке первого слоя, заморозка последующих слоев не привела к существенному увеличению точности распознавания.

Принцип переноса знаний может использоваться для создания многоязычных систем, в этом случае обычно применяется либо многоязычное смешивание (англ. multilingual mix), либо многозадачное обучение. При многоязычном смешивании выполняется обучение единой акустической модели для всех языков (обучение выполняется на всех имеющихся речевых данных) и используется единый набор фонем, включающий в себя все фонемы для всех языков (рисунок 3а). При декодировании речи для каждого языка используется эта единая акустическая модель и свой словарь произношений и модель языка. Многозадачное обучение состоит в обучении одной модели для разных задач. В контексте автоматического распознавания речи, в многозадачном обучении каждый язык рассматривается как задача. Таким образом, скрытые слои ИНС являются общими для всех языков, при этом модель имеет несколько выходных слоев – для каждой модели свой (рисунок 3б). Данные подходы к обучению многоязычной системе были применены,

в частности, в работах [58, 59], где также исследовался перенос знаний, при котором переносились веса скрытых слоев, а веса выходного слоя дообучались для каждого целевого языка на данных этого языка.

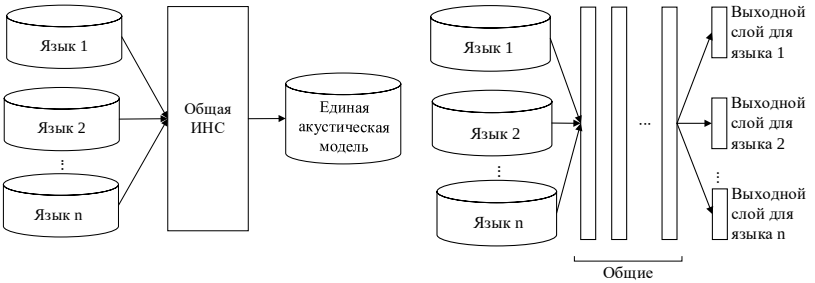

Рис. 3. Основные методы построения многоязычных систем: а) многоязычное смешивание; б) многозадачное обучение

Также многозадачное обучение было применено в работе [60] для ибанского языка, однако в данной работе отдельными задачами являлись распознавание состояний фонем и определение моментов изменения дифференциальных признаков фонем ("landmarks"). Вначале ИНС обучалась на данных английского языка, размеченных на фонемы и "landmarks", а затем выполнялась ее адаптация для малоресурсного ибанского языка на данных, размеченных только на слова, при этом выходные данные детектора "landmarks" для английского языка использовались в качестве справочных меток для обучения акустической модели ибанского языка. Еще одним примером применения многозадачного обучения для малоресурсного языка является [61], где исследуется применение различных типов акустических единиц.

Существуют предварительно обученные многоязычные модели, таких как mBERT или XLM [62-64], в основе которых лежит нейросетевая архитектура трансформер. Эти модели обучены на большом наборе текстов на разных языках и имеют общий для всех языков словарь. Данные модели обучены на примерно 100 языках, и использование их для предварительного обучения малоресурсных языков, не представленных в модели, является недостаточно эффективным. В частности, в работе [65] было показано, что обученная с нуля модель BERT для финского языка превосходит использование mBERT. Однако в работе [66] были проведены

эксперименты по применению предварительно обученных многоязычных моделей для распознавания речи для ряда африканских языков со смешением с английским языком (англ. code-switching), которые показали, что применение таких моделей позволяет снизить WER.

Для сравнения в таблице 2 представлены основные методы переноса знаний, а также относительное снижение ошибки распознавание слов (WER)/символов (CER), полученное различными исследователями за счет применение представленных методов.

Таблица 2. Результаты применения метода переноса знаний для малоресурсных языков

Работа	Нецелевой язык (языки)	Целевой язык	Методы	Относительное сокращение ошибки распознавания слов (WER, %)
[54]	нидерландский	африкаанс	Гетероскедастический линейный дискриминантный анализ и адаптация к речи диктора	34,34‡
[55]	английский + путунхуа	амхарский	Перенос параметров предобученной модели для инициализации параметров целевой модели	37,73
[56]	английский	сенека	Перенос знаний	25,05
			Перенос знаний + аугментация	41,13

‡ В данной работе в качестве метрики используется ошибка распознавания букв (CER, %).

Продолжение Таблицы 2

Работа	Нецелевой язык (языки)	Целевой язык	Методы	Относительное сокращение ошибки распознавания слов (WER, %)
[57]	английский	швейцарский немецкий	Перенос параметров предобученной модели без заморозки параметров	-2,70
			С заморозкой параметров 1 и 2 слоя ИНС	9,46
[58]	амхарский,	Языки из GlobalPhone + речевые данные близкого языка	Переносились веса скрытых слоев, а веса выходного слоя дообучались для каждого целевого языка на данных этого языка	2,61
	тигринья			-0,48
	оромо			0,87
	воламо			1,81
[59]	уйгурский	Языки из GlobalPhone	Перенос параметров многоязычной модели в целевую модель	33,21
[60]	ибан	Языка из TIMIT	Многозадачное обучение	1,90-5,90 (в зависимости от количества обучающих данных)
[66]	зулу	M-BERT, 104 языка	GPT-2 LSTM	5,53
	коса			2,09
	сесото			50,07
	тсвана			41,00

Следует отметить, что конечный результат работы системы распознавания определяется не только примененными авторами методами переноса знаний, а также используемыми наборами обучающих данных и архитектурой ИНС. Несмотря на невозможность прямого сравнения результатов, полученных в работах из таблицы 2, можно сделать вывод о том, что применение методов переноса знаний

позволяет снизить ошибки распознавания, однако итоговый результат сильно зависит от качества и количества обучающих данных, а также примененного классификатора.

7. Заключение. В настоящей статье были рассмотрены основные методы, применяемые при создании систем распознавания речи для малоресурсных языков. Двумя главными способами решения этой проблемы являются расширение обучающих корпусов (аугментация данных) и перенос параметров моделей, обученных на данных других языков, для инициализации параметров модели целевого языка (перенос знаний). В целом, можно констатировать, что оба подхода позволяют достичь определенных результатов, однако каждый из них обладает определенными достоинствами и недостатками.

Методы аугментации могут быть признаны оправданными только в том случае, если для конкретного малоресурсного языка уже существуют какие-то наборы данных. В том случае, если доступные языковые данные чрезвычайно малы, аугментации может быть недостаточно. С другой стороны, при наличии данных, аугментирование языкового материала зачастую оказывается единственным способом создания набора данных для обучения.

Количественное сравнение различных методов переноса знаний затруднено, поскольку, в конечном итоге, результат зависит не только от примененного метода, но и от набора данных. Тем не менее, можно сделать вывод о том, что в том случае, если доступны данные по языку с достаточными ресурсами, имеет смысл работать не с несколькими языками, а с одной языковой парой. В отличие от многозадачных систем, в данном случае осуществляется прямой перенос знаний с одной модели языка на другую. Метод переноса знаний стал широко использоваться в системах автоматического распознавания речи с развитием глубоких ИНС, чаще всего он применяется при разработке интегральных систем распознавания речи. Данный подход особенно эффективен при наличии предобученной модели для языка, близкого целевому, однако даже для сильно отличающихся друг от друга языков данный метод позволяет повысить точность распознавания речи [67]. Для небольших наборов языковых данных модель, использующая многозадачное обучение, может существенно превзойти модель, разработанную для решения одной задачи.

Хотя многоязычные модели сталкиваются с рядом проблем, таких, как использование большого количества разноструктурных языков, дисбаланс данных и другие расхождения, касающиеся таких факторов, как стиль и предметная область [68, 69], тем не менее,

многоязычный подход на сегодня является наиболее перспективным из всех существующих систем: так, одной из наиболее эффективных многоязычных моделей, применяемых к материалу малоресурсных языков, является многоязычный mBART [70]. Также стоит отметить, что многоязычный подход позволяет решить такую проблему, как смешение языков, особенно характерную для многих малоресурсных языков.

В целом, на основании рассмотренных работ можно сделать вывод о том, что основным вектором развития речевых технологий для работы с малоресурсными языками, помимо собственно сбора и аннотации языковых данных, является создание сбалансированных многоязычных систем, пригодных для создания акустических и языковых моделей в рамках конкретных групп близкородственных или структурно схожих языков.

С 2022 года сотрудниками СПб ФИЦ РАН ведутся работы в рамках проекта по созданию системы автоматического распознавания речи на карельском языке (ливвиковское наречие). Несмотря на давнюю литературную традицию и интерес лингвистов к языку и фольклору карел, карельский язык относится к малоресурсным. На момент написания этой статьи основные усилия участников проекта были направлены на сбор и предварительную обработку речевых и текстовых данных, в дальнейшем планируется применить методы аугментации данных, а также использовать метод переноса знаний при обучении моделей.

Литература

1. Magueresse A., Carles V., Heetderks E. Low-resource Languages: A Review of Past Work and Future Challenges // arXiv preprint arXiv:2006.07264. 2020. pp. 1–14.
2. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 6282–6293.
3. Bender E.M. On achieving and evaluating language-independence in NLP. Linguistic Issues in Language Technology. 2011. vol. 6. no. 3. pp. 1–26.
4. Ponti E.M., O'Horan H., Berzak Y., Vulic I., Reichart R., Poibeau T., Shutova E., Korhonen A. Modeling language variation and universals: A survey on typological linguistics for natural language processing // Computational Linguistics. 2019. vol. 45. no. 3. pp. 559–601.
5. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // Proceedings of International workshop on speech and computer (SPECOM-2003). 2003. pp. 8–15.
6. Berment V. Méthodes pour informatiser des langues et des groupes de langues «peu dotées». Doct. Diss. Grenoble, 2004.

7. Cieri Ch., Maxwell M., Strassel S., Tracey J. Selection criteria for low resource language programs // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 4543–4549.
8. Tsvetkov Y. 2017. Opportunities and challenges in working with low-resource languages. Presentation. Carnegie Mellon University, June 22, 2017.
9. Романенко А.Н. Робастное распознавание речи для низко-ресурсных языков. дис. канд. техн. наук: 05.13.11 Ульм. 2020 (на правах рукописи).
10. Мурадова А.Р. Как исчезают языки и как их возрождают // Языковое разнообразие в киберпространстве: российский и зарубежный опыт. Сборник аналитических материалов М.: МЦБС, 2008. С. 70–75.
11. Хусаинов А.Ф., Сулейманов Д.Ш. Система автоматического распознавания речи на татарском языке // Программные продукты и системы. 2013. №4. С. 31–34.
12. Израилова Э.С. О создании фонетико-акустической базы в рамках синтеза чеченской речи // Компьютерная лингвистика и обработка естественного языка. 2017. №2. С. 111–115.
13. Boyko T., Zaitseva N., Krizhanovskaya N., Krizhanovsky A., Novak I., Pellinen N., Rodionova A. The Open corpus of the Veps and Karelian languages: overview and applications // KnE Social Sciences. 2022. vol. 7. no. 3. pp. 29–40.
14. Марковников Н.М., Кипяткова И.С. Аналитический обзор интегральных систем распознавания речи // Труды СПИИРАН. 2018. Вып. 58. С. 77–110.
15. Besacier L., Barnard E., Karpov A., Schultz T. Automatic speech recognition for under-resourced languages: A survey // Speech communication. 2014. vol. 56. pp. 85–100.
16. Карпов А.А., Верходанова В.О. Речевые технологии для малоресурсных языков мира. // Вопросы языкознания. 2015. № 2. С. 117–135.
17. Sabou M., Bontcheva K., Derczynski L., Scharl A. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 859–866.
18. Arora S., Arora K.K., Roy M.K., Agrawal S.S., Murthy B.K. Collaborative speech data acquisition for under resourced languages through crowdsourcing // Procedia Computer Science. 2016. vol. 81. pp. 37–44.
19. Schultz T., Schlippe T. GlobalPhone: Pronunciation Dictionaries in 20 Languages // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 337–341.
20. Strassel S., Tracey J. LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016. pp. 3273–3280.
21. Simpson H., Cieri Ch., Maeda K., Baker K., Onyshkevych B. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources // Collaboration: interoperability between people in the creation of language resources for less-resourced languages. 2008. vol. 7. pp. 7–11.
22. Do V.H., Chen N.F., Lim B.P., Hasegawa-Johnson M.A. Acoustic Modeling for Under-resourced Language using Mismatched Transcriptions // International Journal of Asian Language Processing. 2017. vol. 27. no. 2. pp. 141–153.
23. Hasegawa-Johnson M.A., Jyothi P., McCloy D., Mirbagheri M., Liberto, di G.M. Das A., Ekin B., Liu Ch., Manohar V., Tang H., Lalor E., Chen N.A. Hager P., Kekona T., Sloan R., Lee A.K.C. ASR for under-resourced languages from probabilistic transcription // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017. vol. 25. no. 1. pp. 50–63.

24. Yu C., Kang M., Chen Y., Wu J., Zhao X. Acoustic modeling based on deep learning for low-resource speech recognition: An overview // IEEE Access. 2020. vol. 8. pp. 163829-163843.
25. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition // Proceedings of the 16th Annual Conference of the International Speech Communication Association. 2015. pp. 3586-3589.
26. 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.
27. Park D.S., Chan W., Zhang Y., Chiu C.C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition // Proceedings of Interspeech. 2019. pp. 2613-2617.
28. Hartmann W., Ng T., Hsiao R., Tsakalidis S., Schwartz R. Two-Stage Data Augmentation for Low-Resourced Speech Recognition // Proceedings of Interspeech. 2016. pp. 2378-2382.
29. Jin Z., Finkelstein A., DiVerdi S., Lu J., Mysore G.J. Cute: A concatenative method for voice conversion using exemplar-based unit selection // Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16). 2016. pp. 5660-5664.
30. Goodfellow I.J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair Sh., Courville A., Bengio Y. Generative adversarial nets // Advances in Neural Information Processing Systems 27 / Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (eds.). pp. 2672–2680 // Curran Associates, Inc., 2014.
31. Hsu Ch.-Ch., Hwang H.-T., Wu Y.-Ch., Tsao Y., Wang H. Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks // arXiv preprint arXiv:1704.00849. 2017. pp. 1–5.
32. Kameoka H., Kaneko T., Tanaka K., Hojo N. StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks // Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT'18). 2018. pp. 266–273.
33. Gokay R., Yalcin H. Improving low resource turkish speech recognition with data augmentation and TTS // Proceedings of 2019 16th International Multi-Conference on Systems, Signals and Devices (SSD). 2019. pp. 357–360.
34. Shen J., Pang R., Weiss R.J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhan Y., Wang Y., Skerrv-Ryan R., Saurous R.A., Agiomvrgiannakis Y., Wu Y. Natural TTS Synthesis by Conditioning WaveNet on Mel SpectrogramPredictions // Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 4779–4783.
35. Dua M., Kadyan V., Banthia N., Bansal A., Agarwal T. Spectral warping and data augmentation for low resource language ASR system under mismatched conditions // Applied Acoustics. 2022. vol. 190. 108643.
36. Du C., Yu K. Speaker augmentation for low resource speech recognition // Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 7719-7723.
37. Bagchi D., Wotherspoon Sh., Jiang Zh., Muthukumar P. Speech Synthesis as Augmentation for Low-Resource ASR // arXiv preprint arXiv:2012.13004. 2020. pp. 1–4.
38. Hsu B.J. Generalized linear interpolation of language models // Proceedings of 2007 IEEE workshop on automatic speech recognition & understanding (ASRU). 2007. pp. 136–140.
39. Kurimo M., Enarvi S., Tilk O., Varjokallio M., Mansikkaniemi A., Alumäe T. Modeling under-resourced languages for speech recognition // Language Resources and Evaluation. 2017. vol. 51. no. 4. pp. 961–987.

40. Fadaee M., Bisazza A., Monz Ch. 2017. Data Augmentation for Low-Resource Neural Machine Translation // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. pp. 567–573.
41. 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.
42. Kobayashi S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 2 (Short Papers). 2018. pp. 452–457.
43. Kumar V., Choudhary A., Cho E. Data Augmentation using Pre-trained Transformer Models // Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020. pp. 18–26.
44. Ogawa A., Tawara N., Delcroix M. Language Model Data Augmentation Based on Text Domain Transfer // Proceedings of Interspeech. 2020. pp. 4926–4930.
45. Sennrich R., Haddow B., Birch A. Improving Neural Machine Translation Models with Monolingual Data // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. pp. 86–96.
46. Shahnawazuddin S., Nagaraj A., Kunal K., Aayushi P., Waquar A. Voice Conversion Based Data Augmentation to Improve Children Speech Recognition in Limited Data Scenario // Proceedings of Interspeech 2020. pp. 4382–4386.
47. Tachibana H., Uenoyama K., Aihara S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 4784–4788.
48. Edunov S., Ott M., Auli M., Grangier D. Understanding Back-Translation at Scale // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. pp. 489–500.
49. Fadaee M., Monz Ch. Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. pp. 436–446.
50. Karakanta A., Dehdari J., Genabith, van J. Neural machine translation for low-resource languages without parallel corpora // Machine Translation. 2018. vol. 32. no. 1-2. pp. 167–189.
51. Xia M., Kong X., Anastasopoulos A., Neubig G. Generalized Data Augmentation for Low-Resource Translation // Proceedings of the 57th Annual Meeting of the ACL. 2019. pp. 5786–5796.
52. Wang D., Zheng T.F. Transfer learning for speech and language processing // Proceedings of 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2015. pp. 1225–1237.
53. Gauvain J.-L., Lee C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains // IEEE Transactions on Speech and audio processing. 1994. vol. 2. no. 2. pp. 291–298.
54. 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.
55. Woldemariam Y. Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic // Proceedings 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.

56. Thai B., Jimerson R., Arcoraci D., Prud'hommeaux E., Ptucha R. Synthetic data augmentation for improving low-resource ASR // Proceedings of 2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW). 2019. pp. 1–9.
57. Eberhard O., Zesch T. Effects of Layer Freezing on Transferring a Speech Recognition System to Under-resourced Languages // Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021). 2021. pp. 208–212.
58. Tachbelie M.Y., Abate S.T., Schultz T. Development of Multilingual ASR Using GlobalPhone for Less-Resourced Languages: The Case of Ethiopian Languages // Proceedings of Interspeech. 2020. pp. 1032–1036.
59. Tachbelie M.Y., Abate S.T., Schultz T. Multilingual speech recognition for GlobalPhone languages // Speech Communication. 2022. vol. 140. pp. 71–86.
60. He D., Lim B.P., Yang X., Hasegawa-Johnson M.A., Chen D. Improved ASR for under-resourced languages through multi-task learning with acoustic landmarks // Proceedings of Interspeech. 2018. pp. 2618-2622.
61. Fantaye T.G., Yu J., Hailu T.T. Investigation of Various Hybrid Acoustic Modeling Units via a Multitask Learning and Deep Neural Network Technique for LVCSR of the Low-Resource Language, Amharic // IEEE Access. 2019. T. 7. pp. 105593-105608.
62. Açarçiçek H., Çolakoğlu T., Hatipoğlu P., Huang Ch.H., Peng W. Filtering Noisy Parallel Corpus using Transformers with Proxy Task Learning // Proceedings of the Fifth Conference on Machine Translation. 2020. pp. 940–946.
63. Keung Ph., Salazar J., Lu Y., Smith N.A. Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings // Transactions of the Association for Computational Linguistics. 2020. vol. 8. pp. 828–841.
64. Sun Y., Zhu Sh., Yifan F., Mi Ch. Parallel sentences mining with transfer learning in an unsupervised setting // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 2021. pp. 136–142.
65. Virtanen A., Kanerva J., Ilo R., Luoma J., Luotolahti J., Salakoski T., Ginter F., Pyysalo S. Multilingual is not enough: BERT for Finnish // arXiv preprint arXiv:1912.07076. 2019. pp. 1–14.
66. Vüren, van J.M.J., Niesler T. Improving N-Best Rescoring in Under-Resourced Code-Switched Speech Recognition Using Pretraining and Data Augmentation // Preprints. 2022. 2022050066.
67. Кипяткова И.С., Марковников Н.М. Исследование методов улучшения интегральных систем распознавания речи при недостатке обучающих данных // Труды III Всероссийской акустической конференции. 2020. С. 361-367.
68. Arivazhagan N., Bapna A., Firat O., Lepikhin D., Johnson M., Krikun M., Chen M.X., Cao Y., Foster G., Cherry C., Macherey W., Chen Zh., Wu Y. Massively multilingual neural machine translation in the wild: Findings and challenges // arXiv preprint arXiv:1907.05019. 2019. pp. 1–27.
69. Chathuranga Sh., Ranathunga S. Classification of Code-Mixed Text Using Capsule Networks // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021. pp. 256–263.
70. Stickland A.C., Li X., Ghazvininejad M. Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021. pp. 3440–3453.

Кипяткова Ирина Сергеевна — канд. техн. наук, старший научный сотрудник, лаборатория речевых и многомодальных интерфейсов, 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.: 8(812)328-0421.

Кагиров Ильдар Амирович — научный сотрудник, лаборатория речевых и многомодальных интерфейсов, 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.: 8(812)328-0421.

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

DOI 10.15622/ia.21.4.2

I. KIPYATKOVA, I. KAGIROV

ANALYTICAL REVIEW OF METHODS FOR SOLVING DATA SCARCITY ISSUES REGARDING ELABORATION OF AUTOMATIC SPEECH RECOGNITION SYSTEMS FOR LOW-RESOURCE LANGUAGES

Kipyatkova I., Kagirov I. **Analytical Review of Methods for Solving Data Scarcity Issues Regarding Elaboration of Automatic Speech Recognition Systems for Low-Resource Languages.**

Abstract. In this paper, principal methods for solving training data issues for the so-called low-resource languages are discussed, regarding elaboration of automatic speech recognition systems. The notion of low-resource languages is studied and a working definition is coined on the basis of a number of papers on this topic. The main difficulties associated with the application of classical approaches to automatic speech recognition to the material of low-resource languages are determined, and the principal methods used to solve these problems are outlined. The paper discusses the methods for data augmentation, transfer learning and collection of new language data in detail. Depending on specific tasks, methods for audio material and text data augmentation, transfer learning and multi-task learning are distinguished. In Section 4 of the paper the current information support methods, databases and the basic principles of their architecture are discussed with regard to low-resource languages. Conclusions are drawn about the justification of augmentation and knowledge transfer methods for languages with low information support. In the case of unavailability of language data or structurally similar parent models, the preferred option is to collect a new database, including the crowdsourcing technique. Multilanguage learning models are effective for small datasets. If big language data are available, the most efficient method is transfer learning within a language pair. The conclusions made in the course of this this review will be applied to the data of the low-resource Karelian language, for which an automatic speech recognition system has been being created by the authors of this paper since the beginning of the year 2022.

Keywords: low-resource languages, speech data augmentation, transfer learning, machine learning, language corpora.

References

1. Magueresse A., Carles V., Heetderks E. Low-resource Languages: A Review of Past Work and Future Challenges. arXiv preprint arXiv:2006.07264. 2020. pp. 1–14.
2. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 6282–6293.
3. Bender E.M. On achieving and evaluating language-independence in NLP. Linguistic Issues in Language Technology. 2011. vol. 6. no. 3. pp. 1–26.
4. Ponti E.M., O'Horan H., Berzak Y., Vulic I., Reichart R., Poibeau T., Shutova E., Korhonen A. Modeling language variation and universals: A survey on typological linguistics for natural language processing. Computational Linguistics. 2019. vol. 45. no. 3. pp. 559–601.
5. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. Proceedings of International workshop on speech and computer (SPECOM-2003). 2003. pp. 8–15.
6. Berment V. Méthodes pour informatiser des langues et des groupes de langues «peu dotées». Doct. Diss. Grenoble, 2004.

7. Cieri Ch., Maxwell M., Strassel S., Tracey J. Selection criteria for low resource language programs. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 4543–4549.
8. Tsvetkov Y. 2017. Opportunities and challenges in working with low-resource languages. Presentation. Carnegie Mellon University, June 22, 2017.
9. Romanenko A.N. [Robust Speech Recogniyion for Low-Resource Languages.] Robastnoe raspoznavanie rechi dlja nizko-resursnyh jazykov. Diss. Dokt. Estestvennykh Nauk: 05.13.11. [A doctoral thesis for the academic degree of Dr.rer.nat: 05.13.11] Ulm. 2020 (In Russ.).
10. Muradova A.R. [How languages disappear and how they are revived.] Jazykovoe raznoobrazie v kiberprostranstve: rossijskij i zarubezhnyj opyt. [Linguistic Diversity in Cyberspace: Russian and Foreign Experience]. Moscow: MCBS, 2008. pp. 70–75. (In Russ.)
11. Husainov A.F., Suleimanov D.Sh. [A System for Automatic Tatar Speech Recognition]. Programmnye produkty i sistemy – Program Products and Systems. 2013. №4. pp. 31–34. (In Russ.)
12. Izrailova E.S. [On Elaboration of a Phono-Acoustic Basis within the Framework of the Chechen Speech]. Komp'juternaja lingvistika i obrabotka estestvennogo jazyka – Computer Linguistics and Natural Language Processing. 2017. №2. pp. 111–115. (In Russ.)
13. Boyko T., Zaitseva N., Krizhanovskaya N., Krizhanovsky A., Novak I., Pellinen N., Rodionova A. The Open corpus of the Veps and Karelian languages: overview and applications. KnE Social Sciences. 2022. vol. 7. no. 3. pp. 29–40.
14. Markovnikov N.M., Kipyatkova I.S. [An Analytic Survey of End-to-End Speech Recognition Systems]. Trudy SPIIRAN – SPIIRAS Proceedings. 2018. vol. 58. pp. 77–110.
15. Besacier L., Barnard E., Karpov A., Schultz T. Automatic speech recognition for under-resourced languages: A survey. Speech communication. 2014. vol. 56. pp. 85–100.
16. Karpov A.A., Verkhodanova V.O. [Speech Technologies for Under-Resourced Languages of the World]. Voprosy Jazykoznanija – Problems of Linguistics. 2015. vol. 2. pp. 117–135. (In Russ)
17. Sabou M., Bontcheva K., Derczynski L., Scharl A. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 859–866.
18. Arora S., Arora K.K., Roy M.K., Agrawal S.S., Murthy B.K. Collaborative speech data acquisition for under resourced languages through crowdsourcing. Procedia Computer Science. 2016. vol. 81. pp. 37–44.
19. Schultz T., Schlippe T. GlobalPhone: Pronunciation Dictionaries in 20 Languages. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 337–341.
20. Strassel S., Tracey J. LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016. pp. 3273–3280.
21. Simpson H., Cieri Ch., Maeda K., Baker K., Onyshkevych B. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. Collaboration: interoperability between people in the creation of language resources for less-resourced languages. 2008. vol. 7. pp. 7–11.

22. Do V.H., Chen N.F., Lim B.P., Hasegawa-Johnson M.A. Acoustic Modeling for Under-resourced Language using Mismatched Transcriptions. International Journal of Asian Language Processing. 2017. vol. 27. no. 2. pp. 141–153.
23. Hasegawa-Johnson M.A., Jyothi P., McCloy D., Mirbagheri M., Liberto, di G.M. Das A., Ekin B., Liu Ch., Manohar V., Tang H., Lalor E., Chen N.A. Hager P., Kekona T., Sloan R., Lee A.K.C. ASR for under-resourced languages from probabilistic transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017. vol. 25. no. 1. pp. 50–63.
24. Yu C., Kang M., Chen Y., Wu J., Zhao X. Acoustic modeling based on deep learning for low-resource speech recognition: An overview. IEEE Access. 2020. vol. 8. pp. 163829-163843.
25. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition. Proceedings of the 16th Annual Conference of the International Speech Communication Association. 2015. pp. 3586-3589.
26. 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.
27. Park D.S., Chan W., Zhang Y., Chiu C.C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proceedings of Interspeech. 2019. pp. 2613–2617.
28. Hartmann W., Ng T., Hsiao R., Tsakalidis S., Schwartz R. Two-Stage Data Augmentation for Low-Resourced Speech Recognition. Proceedings of Interspeech. 2016. pp. 2378-2382.
29. Jin Z., Finkelstein A., DiVerdi S., Lu J., Mysore G.J. Cute: A concatenative method for voice conversion using exemplar-based unit selection. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16). 2016. pp. 5660-5664.
30. Goodfellow I.J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair Sh., Courville A., Bengio Y. Generative adversarial nets. Advances in Neural Information Processing Systems 27. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (eds.). pp. 2672–2680. Curran Associates, Inc., 2014.
31. Hsu Ch.-Ch., Hwang H.-T., Wu Y.-Ch., Tsao Y., Wang H. Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks. arXiv preprint arXiv:1704.00849. 2017. pp. 1–5.
32. Kameoka H., Kaneko T., Tanaka K., Hojo N. StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks. Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT'18). 2018. pp. 266–273.
33. Gokay R., Yalcin H. Improving low resource turkish speech recognition with data augmentation and TTS. Proceedings of 2019 16th International Multi-Conference on Systems, Signals and Devices (SSD). 2019. pp. 357–360.
34. Shen J., Pang R., Weiss R.J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhan Y., Wang Y., Skerrv-Ryan R., Saurous R.A., Agiomvrgiannakis Y., Wu Y. Natural TTS Synthesis by Conditioning WaveNet on Mel SpectrogramPredictions. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 4779–4783.
35. Dua M., Kadyan V., Banthia N., Bansal A., Agarwal T. Spectral warping and data augmentation for low resource language ASR system under mismatched conditions. Applied Acoustics. 2022. vol. 190. 108643.
36. Du C., Yu K. Speaker augmentation for low resource speech recognition. Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 7719-7723.

37. Bagchi D., Wotherspoon Sh., Jiang Zh., Muthukumar P. Speech Synthesis as Augmentation for Low-Resource ASR. arXiv preprint arXiv:2012.13004. 2020. pp. 1–4.
38. Hsu B.J. Generalized linear interpolation of language models. Proceedings of 2007 IEEE workshop on automatic speech recognition & understanding (ASRU). 2007. pp. 136–140.
39. Kurimo M., Enarvi S., Tilk O., Varjokallio M., Mansikkaniemi A., Alumäe T. Modeling under-resourced languages for speech recognition. Language Resources and Evaluation. 2017. vol. 51. no. 4. pp. 961–987.
40. Fadaee M., Bisazza A., Monz Ch. 2017. Data Augmentation for Low-Resource Neural Machine Translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. pp. 567–573.
41. 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.
42. Kobayashi S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 2 (Short Papers). 2018. pp. 452–457.
43. Kumar V., Choudhary A., Cho E. Data Augmentation using Pre-trained Transformer Models // Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020. pp. 18–26.
44. Ogawa A., Tawara N., Delcroix M. Language Model Data Augmentation Based on Text Domain Transfer. Proceedings of Interspeech. 2020. pp. 4926–4930.
45. Sennrich R., Haddow B., Birch A. Improving Neural Machine Translation Models with Monolingual Data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. pp. 86–96.
46. Shahnawazuddin S., Nagaraj A., Kunal K., Aayushi P., Waquar A. Voice Conversion Based Data Augmentation to Improve Children Speech Recognition in Limited Data Scenario // Proceedings of Interspeech 2020. pp. 4382–4386.
47. Tachibana H., Uenoyama K., Aihara Sh. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. arXiv preprint arXiv:1710.08969. 2017. pp. 1–5.
48. Edunov S., Ott M., Auli M., Grangier D. Understanding Back-Translation at Scale. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. pp. 489–500.
49. Fadaee M., Monz Ch. Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. pp. 436–446.
50. Karakanta A., Dehdari J., Genabith, van J. Neural machine translation for low-resource languages without parallel corpora. Machine Translation. 2018. vol. 32. no. 1-2. pp. 167–189.
51. Xia M., Kong X., Anastasopoulos A., Neubig G. Generalized Data Augmentation for Low-Resource Translation. Proceedings of the 57th Annual Meeting of the ACL. 2019. pp. 5786–5796.
52. Wang D., Zheng T.F. Transfer learning for speech and language processing. Proceedings of 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2015. pp. 1225–1237.
53. Gauvain J.-L., Lee C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and audio processing. 1994. vol. 2. no. 2. pp. 291–298.

54. 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.
55. Woldemariam Y. Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic. Proceedings 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.
56. Thai B., Jimerson R., Arcoraci D., Prud'hommeaux E., Ptucha R. Synthetic data augmentation for improving low-resource ASR. Proceedings of IEEE Western New York Image and Signal Processing Workshop (WNYISPW). 2019. pp. 1–9.
57. Eberhard O., Zesch T. Effects of Layer Freezing on Transferring a Speech Recognition System to Under-resourced Languages. Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021). 2021. pp. 208–212.
58. Tachbelie M.Y., Abate S.T., Schultz T. Development of Multilingual ASR Using GlobalPhone for Less-Resourced Languages: The Case of Ethiopian Languages. Interspeech. 2020. pp. 1032–1036.
59. Tachbelie M.Y., Abate S.T., Schultz T. Multilingual speech recognition for GlobalPhone languages. Speech Communication. 2022. vol. 140. pp. 71–86.
60. He D., Lim B. P., Yang X., Hasegawa-Johnson M., Chen D. Improved ASR for under-resourced languages through multi-task learning with acoustic landmarks // Proceedings of Interspeech. 2018. pp. 2618–2622.
61. Fantaye T.G., Yu J., Hailu T.T. Investigation of Various Hybrid Acoustic Modeling Units via a Multitask Learning and Deep Neural Network Technique for LVCSR of the Low-Resource Language, Amharic. IEEE Access. 2019. T. 7. pp. 105593-105608.
62. Açarçiçek H., Çolakoğlu T., Hatipoğlu P., Huang Ch.H., Peng W. Filtering Noisy Parallel Corpus using Transformers with Proxy Task Learning. Proceedings of the Fifth Conference on Machine Translation. 2020. pp. 940–946.
63. Keung Ph., Salazar J., Lu Y., Smith N.A. Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings. Transactions of the Association for Computational Linguistics. 2020. vol. 8. pp. 828–841.
64. Sun Y., Zhu Sh., Yifan F., Mi Ch. Parallel sentences mining with transfer learning in an unsupervised setting. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 2021. pp. 136–142.
65. Virtanen A., Kanerva J., Ilo R., Luoma J., Luotolahti J., Salakoski T., Ginter F., Pyysalo S. Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076. 2019. pp. 1–14.
66. Vüren, van J.M.J., Niesler T. Improving N-Best Rescoring in Under-Resourced Code-Switched Speech Recognition Using Pretraining and Data Augmentation. Preprints. 2022. 2022050066.
67. Kipyatkova I.S., Markovnikov N.M. [Investigation of Methods for Improving End-to-End Speech Recognition Systems with a Lack of Training Data.] Trudy III Vserossijskoj akusticheskoj konferencii. [Proceedings of the 3rd All-Russian Acoustic Conference]. 2020. pp. 361-367. (In Russ.)
68. Arivazhagan N., Bapna A., Firat O., Lepikhin D., Johnson M., Krikun M., Chen M.X., Cao Y., Foster G., Cherry C., Macherey W., Chen Zh., Wu Y. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019. 2019. pp. 1–27.
69. Chathuranga Sh., Ranathunga S. Classification of Code-Mixed Text Using Capsule Networks. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021. pp. 256–263.

70. Stickland A.C., Li X., Ghazvininejad M. Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021. pp. 3440–3453.

Kipyatkova Irina — Ph.D., 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: 8(812)328-0421.

Kagirov Ildar — Researcher, 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: 8(812)328-0421.

Acknowledgements. This research is supported by the Russian Science Foundation (project № 22-21-00843).