ГОСУДАРСТВЕННЫЙ КОМИТЕТ ПО НАУКЕ И ТЕХНОЛОГИЯМ ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «ИНСТИТУТ ПРОБЛЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА»

На правах рукописи

Пикалёв Ярослав Сергеевич

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ И ПРОГРАММНЫХ СРЕДСТВ РАСПОЗНАВАНИЯ СЛИТНОЙ РУССКОЙ РЕЧИ

Специальность 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) (технические науки)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Донецк – 2021

Работа выполнена в ГОСУДАРСТВЕННОМ УЧРЕЖДЕНИИ «ИНСТИТУТ ПРОБЛЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА» ГОСУДАРСТВЕННОГО КОМИТЕТА ПО НАУКЕ ТЕХНОЛОГИЯМ ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ, г. Донецк

Научный руководитель: кандидат технических наук, доцент
Ермоленко Татьяна Владимировна
ГОУ ВПО «ДОННУ» (г. Донецк), доцент кафедры компьютерных технологий

Официальные оппоненты: **Харламов Александр Александрович**
доктор технических наук,
ФГБУН «Институт высшей нервной деятельности и нейрофизиологии» (г. Москва), старший научный сотрудник

Бурлаева Екатерина Игоревна
кандидат технических наук,
ГУП ДНР «ЭНЕРГИЯ ДОНБАССА» (г. Донецк), специалист 1 категории отдела сетевых сервисов дирекции по информационным технологиям

Ведущая организация: **Государственное учреждение «Институт прикладной математики и механики (ГУ «ИПММ»)** (г. Донецк)

Защита состоится «12» октября 2021 г. в 14.00 часов на заседании диссертационного совета Д 01.024.04 при ГОУВПО «ДОННТУ» и ГОУВПО «ДОННУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 1, ауд. 203 Тел./факс: 380(62) 304-30-55, e-mail: uchensovet@donntu.org.

С диссертацией можно ознакомиться в библиотеке ГОУВПО «ДОННТУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 2. Адрес сайта университета: http://donntu.org

Автореферат разослан «__» _____ 2021 г.

Ученый секретарь
диссертационного совета Д 01.024.04
кандидат технических наук, доцент Т.В. Завадская

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Автоматическое распознавание речи является динамично развивающимся направлением в области искусственного интеллекта. Задача распознавания речи получила широкое распространение вследствие высокой применимости на практике. Однако, на сегодняшний день в сфере распознавания русскоязычной слитной речи успехи достигнуты только в пределах словарного запаса, связанного с узкой предметной областью, а распознавание слитной речи до сих пор не имеет четкого решения в силу ряда возникающих трудностей, связанных с отсутствием объёмного аннотированного речевого и нормализованного текстового корпуса, необходимого для статистического моделирования языка; флективностью, а также свободным порядком слов во фразе; орфоэпическими нормами, увеличивающими акустическую вариативность русской речи.

В связи с этим, задача совершенствования методов и программных средств дикторонезависимого распознавания слитной русской речи, позволяющих обеспечивать высокое качество распознавания, учитывать особенности русской речи и адаптироваться под любую предметную область, является актуальной и имеет отраслевое значение.

Связь работы с научными программами, планами, темами. В основу диссертационного исследования положены работы, выполненные в Институте проблем искусственного интеллекта в рамках научно-исследовательских работ: «Разработка методов распознавания слитно произнесённых фраз в рамках концепции пофонемного распознавания речи с обобщённой транскрипцией» (№Г/Р 0113U001326); «Исследование и разработка методов семантического анализа и интерпретации потоков данных интеллектуальными системами» (№Г/Р 0118D000003).

Степень разработанности темы исследования. Системы распознавания слитной русскоязычной речи от компаний Google и Яндекс демонстрируют высокую точность распознавания речи – около 75–90%, а методы и модели распознавания для русского языка, как правило, заимствуются из другого языка. Поэтому качество их работы значительно падает при распознавании разговорной речи.

Среди зарубежных исследователей, занимающихся данным направлением, стоит выделить D. Povey, G. Hinton, P. Cosi, A. Graves, O. Abdel-Hamid, A. Baevski, D. Amodei, G. Saon, L. Deng, A. Senior, T. Sainath. Среди российских исследователей следует отметить работы А. Карпова, А. Ронжина, И. Кипятковой, И. Меденникова, Д. Кушнира, И. Тампеля.

Российской компанией ООО «ЦРТ» разработана система автоматической генерации субтитров в режиме реального времени, которая использует искусственные нейросети (ИНС) и обучена на 32 часах записей новостей. Статистическая модель языка построена для конкретной тематики

телепередачи. В результате качество распознавания системы при переходе на другие предметные области сильно снижается.

В рамках российской разработки RealSpeaker применяется видеорасширение для увеличения точности программ распознавания речи, однако данное решение не распознает большинство союзов, а при равномерной диктовке пропадают части фраз.

Таким образом, на настоящий момент не существует систем распознавания слитной русской речи, сопоставимых по качеству с вышеупомянутыми системами для английского языка.

Учитывая вышеизложенное, можно сделать вывод о необходимости повышения эффективности существующих моделей, методов и программных средств для построения систем дикторонезависимого распознавания слитной русской речи, способных адаптироваться под любую предметную область.

Цель диссертационного исследования – повышение эффективности дикторонезависимой системы автоматического распознавания слитной русской речи за счет модернизации алгоритмов системного анализа, обработки и распознавания речевой информации, а также автоматической обработки текстовых данных.

Для достижения цели в работе решены следующие **задачи**:

1) осуществлен выбор методов построения акустической (АМ) и языковой моделей (ЯМ), транскриптора, классификатора фонем и декодера, исходя из анализа современных технологий распознавания слитной русской речи;

2) собраны и обработаны речевые и текстовые данные, находящиеся в открытом доступе, с целью создания аннотированного речевого корпуса для обучения АМ и ЯМ;

3) разработаны методы автоматического построения словаря транскрипций, позволяющие определять позицию ударения, генерировать транскрипцию для слов-исключений и осуществлять практическую транслитерацию с учетом орфоэпических норм русского языка;

4) разработаны методы получения робастных акустических признаков и обучения АМ на основе глубоких нейронных сетей;

5) разработан классификатор для распознавания фонем;

6) оценено качество построенных АМ и классификатора фонем с целью обоснования предложенного подхода их использования в системах дикторонезависимого распознавания слитной русской речи;

7) построена система автоматического распознавания слитной русской речи (Automatic Speech Recognition, ASR) на основе предложенных методов и моделей, проведена оценка качества ее работы по сравнению с российскими и зарубежными системами.

Объект исследования – процессы анализа, обработки и классификации речевого сигнала в системах автоматического распознавания речи.

Предмет исследования – методы и алгоритмы построения АМ и ЯМ, методы распознавания речевого сигнала.

Научная новизна полученных результатов заключается в следующем:

1) получили дальнейшее развитие нейросетевые методы автоматического определения позиции ударения в слове за счет модернизации архитектуры нейросети типа Transformer, которая заключается в увеличении количества слоёв, использовании методов градиентного отсечения и teacher forcing для оптимизации параметра скорости обучения, что позволило повысить точность определения позиции ударения на 10% по сравнению со стандартной моделью Transformer;

2) усовершенствована seq2seq модель для генерации практических транскрипций англоязычных слов и слов-исключений за счет применения механизма обучения с подкреплением и метода beam-search для выбора наиболее вероятной последовательности символов, что позволило повысить точность модели по критерию количества ошибочно сгенерированных символов на 0,8% и 3%, по критерию неправильно сгенерированных слов на 0,6% и 9% соответственно.

3) предложена модель нейросетевой параметризации, основанная на объединении ансамбля нейронных сетей с «узким горлом» и архитектуры ResNet-50. Использование данной модели позволяет повысить точность распознавания на 2,7% по сравнению с моделью, извлекающей стандартные bottleneck-признаки;

4) получили дальнейшее развитие методы нейросетевой классификации фонем за счет использования механизма внимания в последнем скрытом слое сети, включающей в себя нейросеть с временными задержками и двунаправленную нейросеть с долгой кратковременной памятью, что позволило сохранять высокую точность на относительно небольшом обучающем наборе аудиоданных, свойственную системам, для обучения которых требуется речевая база длительностью в десятки тысяч часов.

Теоретическая значимость научных результатов, полученных в ходе диссертационного исследования, определяется созданием нового подхода нейросетевой параметризации речевого сигнала для обучения АМ. В частности, предложенная технология извлечения робастных признаков из скрытых слоев иерархической мультимодульной ИНС вносит свой вклад в теорию понимания кодирования сигнала на основе глубоких нейросетей, которые на сегодняшний день рассматриваются как «черный ящик».

Практическое значение работы. Материалы исследований могут быть использованы при разработке методов автоматического формирования аннотированных речевых баз данных; методов автоматического построения словаря транскрипций системах синтеза речи; методов получения робастных акустических признаков и обучения АМ, а также при разработке классификатора для распознавания фонем в системах голосового управления и поиска по голосовому запросу, а также в системах диктовки с приемлемым уровнем ошибок.

Результаты и выводы диссертационной работы нашли применение в Институте проблем искусственного интеллекта, что подтверждается справкой о внедрении (справка №347/01-01 от 01.12.2020).

Методология и методы исследования. Для решения поставленных задач использовались следующие методы:

– прикладной лингвистики для анализа закономерностей синтаксиса, морфологии и фонетического состава русского языка, анализа структуры существующих словарей;

– методы математической статистики для оценки эффективности разработанных моделей;

– методы цифровой обработки сигналов для получения акустических характеристик речевых сигналов;

– методы машинного обучения для построения АМ и ЯМ и классификации фонем.

Положения, выносимые на защиту.

1. Доказано, что модификация seq2seq модели для генерации практических транскрипций англоязычных слов и слов-исключений на базе архитектуры Transformer за счет применения механизма обучения с подкреплением обеспечивает повышение точности генерации транскрипции по критерию количества ошибочно сгенерированных символов и по критерию неправильно сгенерированных слов.

2. Установлено, что усовершенствование метода акустического моделирования за счет аугментации и модификации признаков для получения адаптивных и дискриминативных характеристик повышает робастность акустических признаков, обеспечивая тем самым их инвариантность к смене диктора и акустической обстановке и повышение точности распознавания.

3. Доказано, что предложенный метод нейросетевой параметризации речевого сигнала на основе иерархической мультимодульной архитектуры MultiBN и архитектуры ResNet-50 позволяет извлечь из скрытых слоев информативные акустические признаки, устойчивые по отношению к темпу речи, акустической среде и междикторской вариативности, что приводит к повышению точности распознавания на по сравнению с моделью, извлекающей стандартные bottleneck-признаки.

По направлению исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) (технические науки) по областям исследований: п.5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации»; п.12 «Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации».

Степень достоверности и апробация результатов обеспечивается полнотой анализа теоретических и практических исследований, положительной оценкой на научных конференциях и семинарах, выполненными публикациями.

Апробация результатов работы. Основные научные положения и результаты диссертационной работы доложены, обговорены и приняты на конференциях: «Искусственный интеллект: теоретические аспекты и практическое применение» (г. Донецк, 2020); Международная научная конференция студентов и молодых учёных «Донецкие чтения» (г. Донецк, 2019, 2018, 2017); Международная научно-техническая конференция «Информатика, управляющие системы, математическое и компьютерное моделирование» (г. Донецк, 2019, 2016); XII Мультиконференция по проблемам управления (г. Геленджик, Дивноморское, 2019); Международная научно-техническая конференция «Интеллектуальные технологии и проблемы математического моделирования» (г. Геленджик, Дивноморское, 2018); VIII Международная конференция по когнитивной науке (г. Светлогорск, 2018).

Личный вклад автора. Соискателем лично решены задачи диссертации. В работах, опубликованных в соавторстве, личный вклад автора заключается в выполнении аналитических расчётов, практических экспериментов, реализации программных решений и статистическом анализе полученных результатов. Все выносимые на защиту положения получены автором лично.

Публикации. Основные научные результаты диссертации опубликованы в 17 научных работах, в том числе в 5 научных статьях в изданиях, рекомендуемых ВАК для публикации трудов на соискание ученых степеней.

Объем и структура диссертации. Диссертация изложена на 180 страницах машинописного текста и состоит из списка сокращений, введения, пяти глав, выводов, заключения, списка литературы, 3 приложений. Работа иллюстрирована 47 рисунками, содержит 11 таблиц. Список литературы включает 186 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационного исследования, сформулированы цель и задачи исследования, показаны научная новизна, значимость работы, представлены положения, выносимые на защиту.

В **первой главе** работы приведен обзор современных технологий распознавания слитной речи.

Рассмотрены технологии извлечения акустических признаков, в качестве которых современными системами распознавания используются MFCC, FBANK и PLP. Приведен обзор техник для модификации извлечённых акустических признаков с целью получения шумоустойчивых, а также адаптивных и дискриминативных характеристик.

Рассмотрены методы акустического моделирования на основе скрытых марковских моделей (Hidden Markov Models, HMM) и гауссовых смешанных моделей (Gaussian Mixture Models, GMM). Применяются преимущественно

смеси с диагональной матрицей ковариации, что влечет за собой необходимость использования некоррелированных признаков.

Перспективным для акустического моделирования представляется нейросетевая параметризация речевого сигнала, т.к. с ростом числа слоёв признаки становятся устойчивее по отношению к темпу речи, акустической среде и междикторской вариативности.

Описано решение задачи языкового моделирования на основе n-gram и различных нейросетевых архитектур.

Для разработки транскриптора перспективно использовать модель Transformer со слоем внимания, которая хорошо параллелизуется и распределяет внимание между частями последовательности.

Приведено описание процесса декодирования на основе конечных автоматов, что позволяет объединить различные источники знаний – АМ и ЯМ, лексиконы.

На основе проведенного анализа сформулированы цель и задачи исследования, показана необходимость совершенствования методов и программных средств дикторонезависимого распознавания слитной русской речи.

Во **второй главе** описаны техники обработки данных для создания речевого корпуса и обучения АМ и ЯМ.

Для проверки соответствия текстовых расшифровок и аудио модифицирован алгоритм Смита-Уотермана, использующийся для нахождения локальных паттернов с высоким уровнем сходства, который запоминает позицию начала и конца совпадения в исходных данных и ищет n максимальных совпадений в случае наличия нескольких вхождений с одинаковыми максимальными оценками.

Для оценки эффективности предложенной модификации использовался показатель пословной ошибки распознавания (Word Error Rate, WER):

$$\mathrm{W}ER = \frac{S+I+D}{N} \cdot 100\ \%,$$

где N – количество слов в тексте; S, I, D – число замен, вставок и удалений в результате распознавания. Предложенная модификация позволила улучшить точность в среднем на 12% относительно базовой реализации.

Для увеличения объема обучающих данных и повышения робастности АМ предложена техника аугментации речевых данных путем наложения шумов и голоса другого диктора, что позволило уменьшить WER на 1,14% для трифонной и на 1,7% для монофонной GMM-HMM моделей.

Предложена техника автоматической нормализации текста для определения языка, расшифровки цифро-буквенных комплексов и сокращений, получения практической транскрипции.

Для определения языка текста использовалась сверточная нейросеть (Рисунок 1) с точностью на тестовой выборке 82,5%.

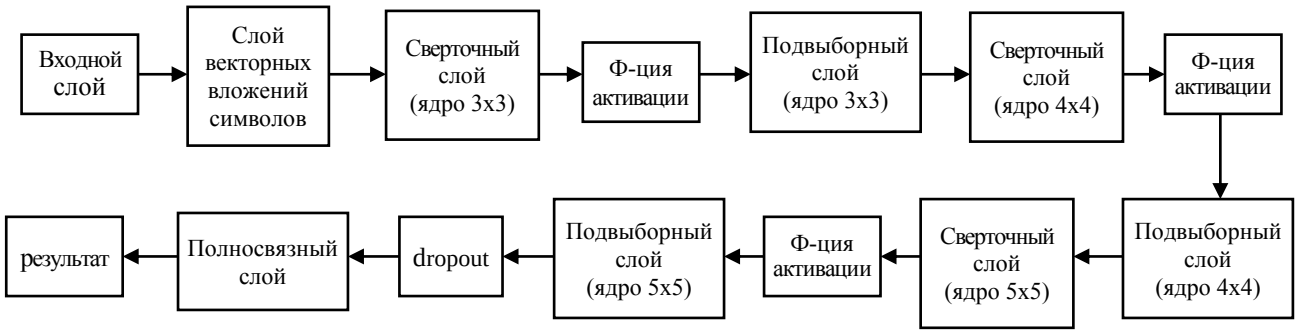

Рисунок 1 – Архитектура нейросети для определения языка текста

Для согласования чисел при расшифровке цифро-буквенных комплексов разработан синтаксический анализатор с применением глубоких ИНС. Используемая модель состоит из трех свёрточных ИНС (200 слоёв в глубину) со слоем внимания. Данные, подающиеся на вход, представлены вектором размерностью в 64 единиц.

Предложенная архитектура формирует контекстный вектор для слов в предложении и позволяет добиться улучшения точности анализа: по критерию Unlabelled Attachment Score (UAS) – 94,3%, по критерию Labelled Attachment Score (LAS) – 90,2%.

Для классификации ошибок предложено векторное представление слов, основанное на архитектуре ULMfit, учитывающей контекст слова. Выходы предпоследнего слоя ULMfit формируют 100-мерный вектор, поступающий на вход ИНС для классификации ошибок с архитектурой QRNN со слоем внимания, что дало точность классификации ошибок на тестовой выборке 96,5%.

Для исправления текстовых ошибок предложена архитектура, использующая модель seq2seq, а в качестве входных данных – последовательность символов, а не слов, что позволяет избежать проблем с внесловарными словами. В качестве слоя векторных вложений символов используется свёрточный слой, извлечённый из предобученной модели векторных представлений языковой модели ELMo, которая генерирует векторное представление для слова на основе его контекста. Нейросеть улучшена при помощи техник teacher forcing и градиентного отсечения. Предложенная архитектура работает быстрее, чем стандартно используемые в NLP-задачах LSTM, и обеспечивает точность более 96%.

В **третьей главе** рассмотрены особенности фонетики русского языка и предложен метод автоматического построения словаря транскрипций.

Авторская система автоматической генерации транскрипций Cyr2Trans использует словари транскрипций и нейросети (Рисунок 2).

Нейронные сети применяются для определения положения ударения (DetAccentNN), генерации транскрипций для слов-исключений (PhonExcNN), а также для практической транслитерации (PractTransNN).

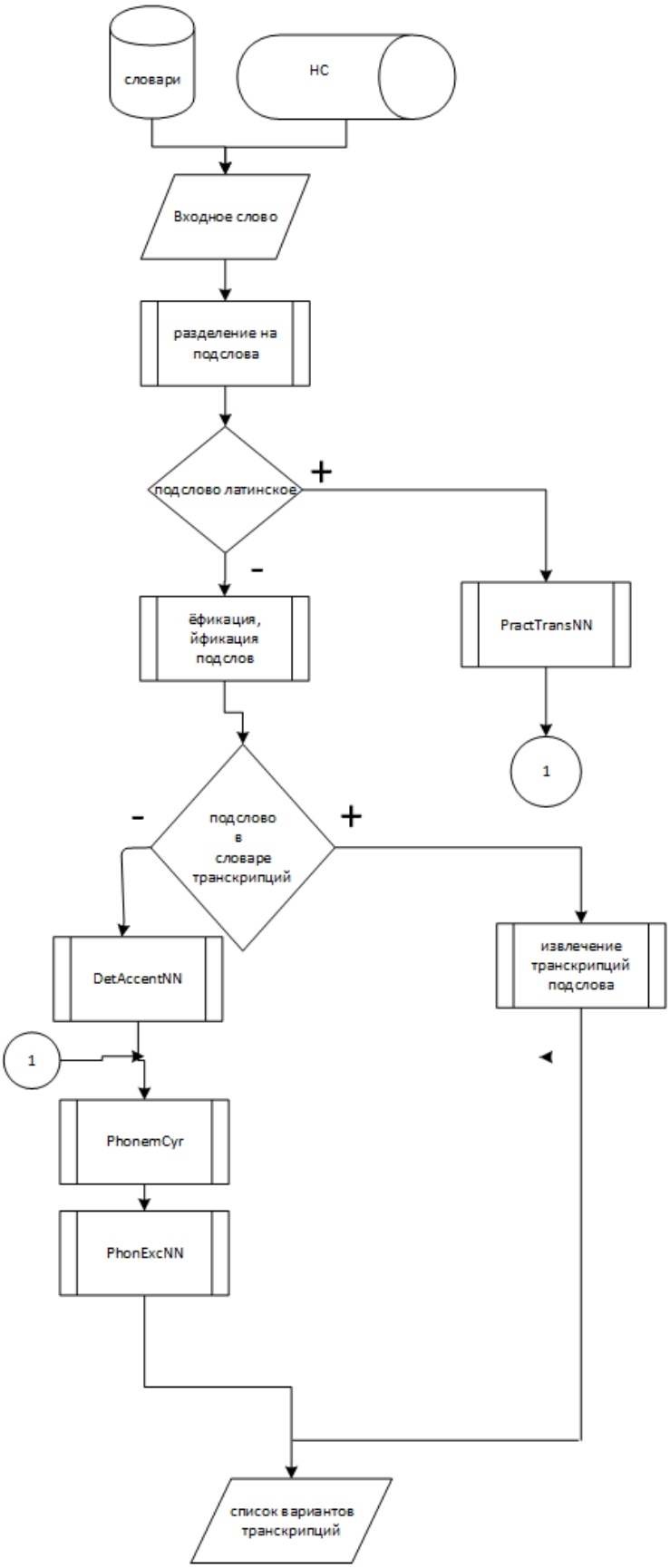

Рисунок 2 – Блок-схема работы системы автоматической генерации транскрипций Cyr2Trans

Для разделения слов на подслова используется модифицированный алгоритм SnowballStemmer. Отличие предложенного алгоритма стемматизации состоит в использовании отдельных словарей суффиксов, постфиксов, окончаний для каждой самостоятельной части речи и общего словаря приставок.

На основе ИНС с архитектурой Transformer автором разработана нейросеть DetAccentNN, использующая техники градиентного отсечения и увеличения количества блоков в энкодере. Предложенная модификация позволила снизить показатель WER до 0,08 и повысить точность генерации транскрипции на 10% по сравнению со стандартным подходом.

Для получения практической транскрипции используется нейросетевой подход и словарь. Предложена нейросетевая модель на архитектуре Transformer, улучшенной при помощи техник teacher forcing, градиентного отсечения и метода beam-search с совместным применением обучения с учителем и обучения с подкреплением, реализованным в RL-block, показанная на рисунке 3, где forward-transformer – ИНС генерации транскрипции, обученная на парах x-y (слово-транскрипция); backward-transformer – ИНС генерации транскрипций, обученная на парах y-x; RL-block – механизм обучения с подкреплением; forward_RL-transformer – итоговая ИНС генерации транскрипции.

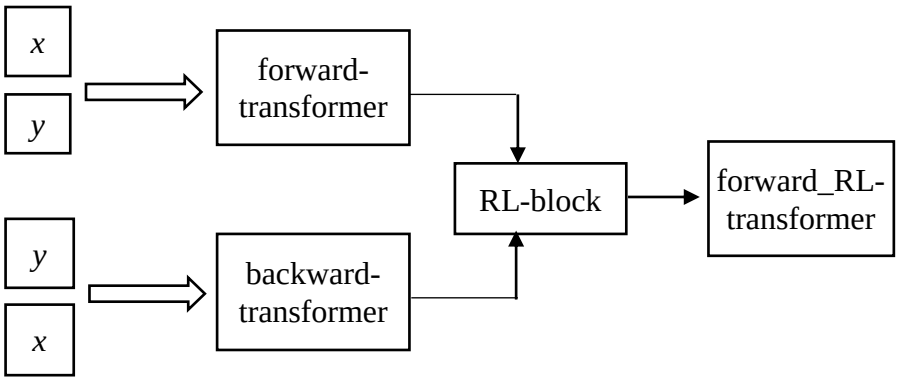

Рисунок 3 – Схема обучения модели PractTransNN

Предложенный подход с использованием глубокого обучения позволяет получать англо-русскую практическую транскрипцию с точностью более 90% для слов и более 95% для символов.

Для получения транскрипции слов-исключений обучена ИНС на наборе данных, состоящего из слов, отличающихся от фонетических норм русского языка, объемом около 10 тыс.

Предложенная техника модернизации моделей типа seq2seq за счет применения механизма обучения с подкреплением позволила повысить точность обученной модели генерации транскрипций для слов-исключений по критерию PER на 9%, по критерию WER – на 3%.

Четвёртая глава посвящена описанию технологии повышения робастности АМ и архитектуры сети для классификации фонем.

Для повышения робастности использовалась аугментация и нейросетевая параметризация речевого сигнала (подход bottleneck), а также техники модификации акустических признаков для получения адаптивных и дискриминативных характеристик.

В качестве акустических признаков используются мел-частотные кепстральные коэффициенты (MFCC), их первые и вторые производные, логарифмы энергии спектра набора треугольных Mel-фильтров (FBANK) и коэффициенты перцептивного линейного предсказания (PLP). Размерность вектора признаков для обучения модели на основе скрытых марковских моделей и гауссовых смесей (HMM-GMM) составляет 43 (40 MFCC, 3 PLP).

Для нейросетевой параметризации обучена иерархическая мультимодульная ИНС MultiBN, на вход которой подаются 100-мерные вектора по 16 тыс. фреймов, данные вектора получаются путём объединения FBANK-признаков и i-векторов при помощи LDA. MultiBN – ансамбль из bottleneck нейросетей, обученных на основе дискриминативного критерия MPE (Рисунок 4).

Каждая из bottleneck-нейросетей состоит из трех скрытых слоёв, по 2048 нейронов в каждом слое. На каждом уровне bottleneck-нейросети производится процедура получения трансформированных весовых коэффициентов (fine-tune), которые и являются нашими информативными признаками.

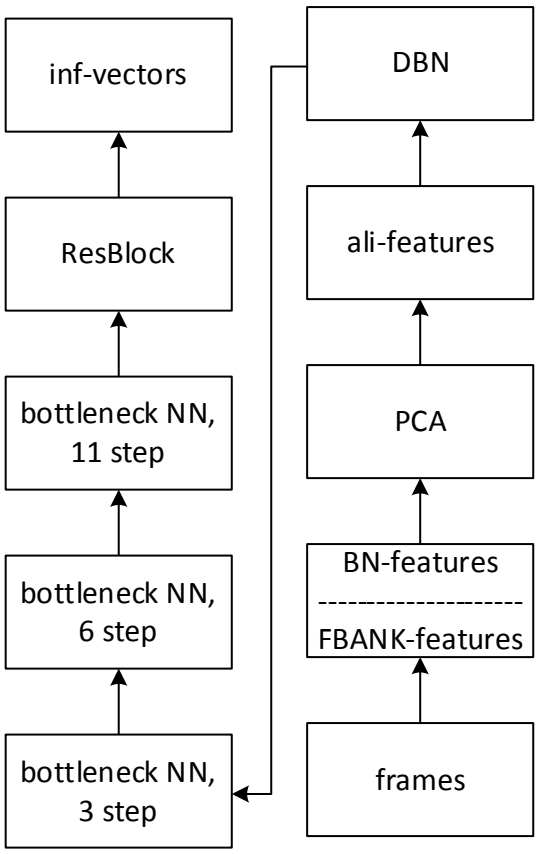

Рисунок 4 – Архитектура нейросети MultiBN

Для формирования акустических признаков первого уровня используется контекст векторов в 2 кадра; для второго уровня – в 5 кадров, а для третьего уровня – в 10 кадров. Признаки третьего уровня подаются на ResBlock, представляющую собой ИНС, основанную на архитектуре ResNet-50, с двумя дополнительными линейными слоями. Архитектуру этой сети демонстрирует рисунок 5, где F – размер входных признаков; D – количество фреймов, N – размер новых признаков; input_features – входные признаки; LinearLayer – линейные нейронные слои; AveragePoolingLayer – слой, трансформирующий данные при помощи свёртки 1*1; FalttenOp – операция трансформации в одномерное пространство; out_emb – выходные информативные признаки.

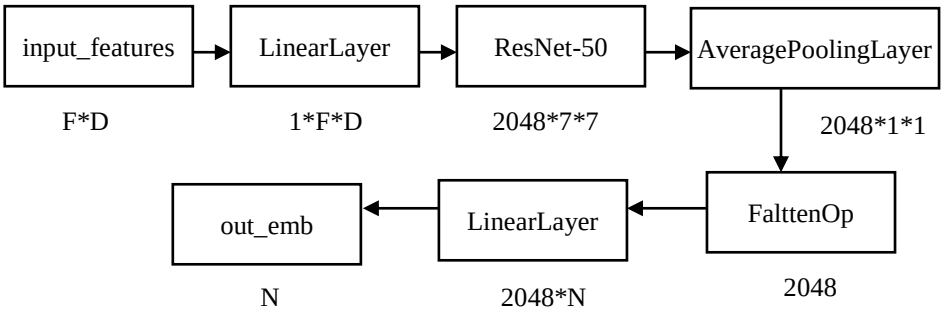

Рисунок 5 – Архитектура ResBlock

Извлеченные из ResBlock признаки объединяются с FBANK при помощи LDA и поступают на вход сети, классифицирующей фонемы.

Нейросетевая модель для предсказания последовательности фонем (Рисунок 6) основана на ИНС с временной задержкой (Time Delay Neural Network, TDNN) и двунаправленной долгой кратковременной нейронной сети (Bidirectional Long Short Memory, BLSTM) со слоем внимания (attention), каждая из них имеет 5 слоёв с 2048 нейронами (1 входной, 1 выходной и 3 скрытых слоя).

Для сопоставления векторов признаков фонемам использовались монофонная и квифонная модели HMM-GMM, для обучения которых последовательно применялись: линейное преобразование признаков, максимизирующее среднее правдоподобие (MLLT), обеспечивающее дикторонезависимость и робастность; аффинное преобразование функции максимизации вероятности как пространство характеристик максимального правдоподобия линейной регрессии (fMLLR) и адаптивное обучение диктора (Speaker Adaptive Training, SAT) для борьбы с внедикторскими вариациями; обучение подпространства моделей гауссовых смесей (SGMM); добавление i-векторов размерностью 100 к вектору признаков для адаптации к диктору и акустической обстановке. Для уменьшения признакового пространства использовался линейный дискриминантный анализ (LDA).

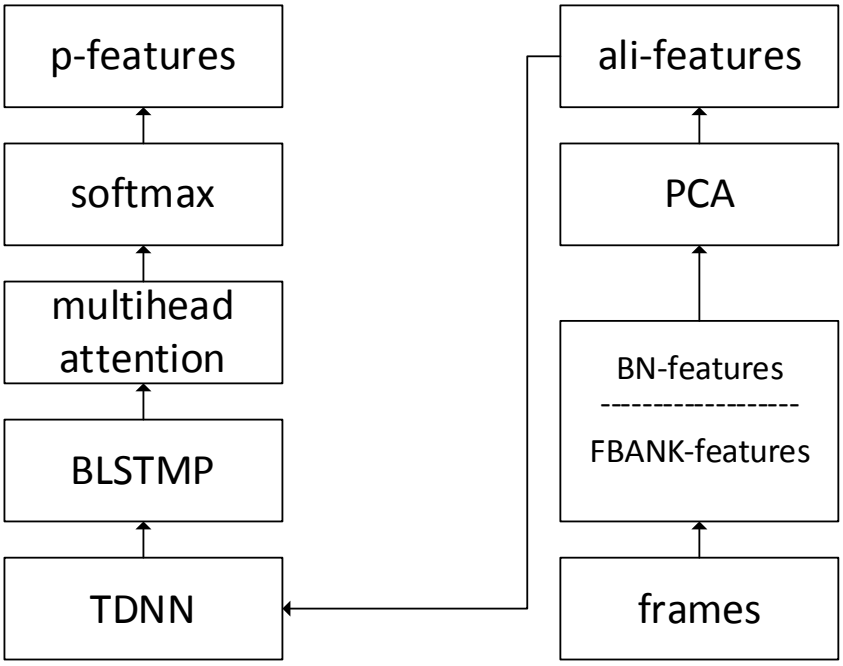

Рисунок 6 – Архитектура TDNN-BLSTM-attention

Для обучения АМ применялся речевой корпус общей продолжительностью порядка 20 часов. ЯМ обучена на основе триграмм, объем словаря – 500 тыс. слов. Обучающая и тестовая выборки с аудиоданными разделены в отношении 95/5. В таблице 1 приведена оценка качества моделей по показателю WER.

Таблица 1 – Сравнение эффективности различных АМ*

№	Описание модели	WER, %	кол-во GMM	кол-во HMM
1	Монофонная HMM-GMM	64.01	4000	1500
2	Квифонная HMM-GMM	35.76	20000	2500
3	Квифонная с применением LDA и MLLT	21.55	50000	4000
4	Модель 3 с применением техники переопределения вероятности тишины	19.98	50000	4000
5	Квифонная с применением SAT и fMLLR	14.07	100000	5000
6	Квифонная, с применением SGMM	10.81	120000	8500
7	Квифонная с применением TDNN-BLSTM-attention и стандартных bottleneck-признаков, объединённых с FBANK	7.93	-	-
8	Квифонная с применением TDNN-BLSTM-attention и объединения акустических признаков FBANK и признаков, извлечённых из ResBlock	**5.2**	-	-

* кол-во GMM – количество гауссиан, кол-во HMM – количество состояний для скрытой марковской модели

Ошибка распознавания Google Cloud Speech API на тех же данных составила 9,33%.

Предложенная модель нейросетевой параметризации позволяет повысить точность распознавания на 2,7% по сравнению с моделью, извлекающей стандартные bottleneck-признаки, а разработанные АМ и архитектура ИНС показывают результат распознавания фонем на 4,1% лучше, чем ASR от Google, которая обучалась на речевых базах объемом в десятки тысяч часов.

В **пятой главе** приведено описание структуры разработанной системы распознавания русской речи. Описан процесс получения ЯМ и АМ, процесс декодирования. Проведено сравнение эффективности работы разработанной системы с существующими решениями.

Для построения ЯМ использовалась ИНС с архитектурой LSTM. Для декодирования применялся взвешенный преобразователь с конечным числом состояний (WFST).

Для сравнения разработанной ASR (ASR_work) с известными аналогами выбрана система компании ЦРТ (CST ASR), а также Google Cloud ASR (Таблица 2). В качестве тестового корпуса использовался набор данных телефонных звонков Russian Open Speech Dataset объемом около 7,2 Гб. В качестве метрик для оценивания систем используются WER, SER и метрика оценки скорости получения результата распознавания (speed rate, SR):

$$SR = \frac{T_{rec}}{T},$$

где T_{rec} – время распознавания сигнала продолжительностью T.

ASR_work оказалась лучше системы CST по качеству распознавания на 10,42%. Google Cloud оказалась лучше ASR_work на 6,44%, но при этом авторская система превосходит остальные по скорости распознавания.

Таблица 2 – Сравнение качества работы ASR-систем

ASR	WER	SER	SR
CST ASR	0,3805	0,96	1,17
Google Cloud	0,2119	0,82	1,05
ASR_work	0,2763	0,91	0,15

ASR_work, обладая достаточной точностью в задаче распознавания слитной русской речи, использует для своего обучения значительно меньше данных и превосходит рассмотренные системы по показателю SR более, чем в 7 раз.

ЗАКЛЮЧЕНИЕ

Диссертация является законченной научно-исследовательской работой, в которой получено решение актуальной научно-технической задачи повышения эффективности дикторонезависимой системы автоматического распознавания слитной русской речи, учитывающей её особенности и адаптирующейся под

любую предметную область за счет модернизации алгоритмов системного анализа, обработки и распознавания речевой информации, а также автоматической обработки текстовых данных. Основные научные результаты и выводы состоят в следующем.

1. Анализ состояния исследований в области распознавания речи показал, что для построения АМ наиболее перспективным представляется нейросетевая параметризация речевого сигнала и модификация акустических признаков для получения адаптивных и дискриминативных характеристик; для построения ЯМ – нейросети с архитектурой LSTM и Transformer; для построения транскриптора – ИНС с архитектурой seq2seq; для пофонемного распознавания – глубокие нейросети; для декодирования – подходы, основанные на WFST-графе.

2. Для формирования собственного аннотированного речевого корпуса:

– модифицирован классический алгоритм парного выравнивания Смита-Уотермана для проверки соответствия текстовых расшифровок и аудио за счет запоминания начала и конца совпадения в исходных данных, что повысило его точность в среднем на 10,5 % по сравнению с исходным алгоритмом;

– предложена техника аугментации речевых данных, позволяющая повысить робастность АМ и уменьшить WER на 1,14% для трифонной АМ и на 1,7% для монофонной;

– разработаны алгоритмы нормализации текстов на основе сверточных сетей, позволяющие: определить язык отдельного предложения с точностью 82,5%; провести согласование чисел для корректной расшифровки цифро-буквенных комплексов с точностью по критерию UAS – 94,3%, по критерию LAS – 90,2%.

В результате использования предложенных алгоритмов для обучения АМ подготовлен речевой корпус длительностью более 29 часов, для обучения ЯМ сформирована база нормализованных текстов объемом 15,2 Гб.

3. Разработаны архитектуры ИНС для автоматического формирования словаря транскрипций. Предложенные нейросетевые модели позволяют:

– автоматически определять позицию ударения в слове за счет модернизации архитектуры ИНС Transformer, которая заключается в увеличении количества слоёв, использовании методов градиентного отсечения для оптимизации параметра скорости обучения, что увеличило точность определения позиции ударения на 10%;

– генерировать практические транскрипции англоязычных слов и слов-исключений путем усовершенствования seq2seq модели за счет применения механизма обучения с подкреплением и метода beam-search для выбора наиболее вероятной последовательности символов, что увеличило точность модели по критерию количества ошибочно сгенерированных символов на 0,8% и 3%, по критерию неправильно сгенерированных слов на 0,6% и 9% соответственно.

4. Для получения робастных акустических признаков и обучения АМ предложена нейросетевая параметризация, основанная на объединении ансамбля нейронных сетей с узким горлом и архитектуры ResNet-50, что позволяет повысить точность распознавания на 2,7% по сравнению с моделью, извлекающей стандартные bottleneck-признаки.

5. Для разработки классификатора фонем усовершенствованы методы нейросетевой классификации за счет использования механизма внимания в последнем скрытом слое сети, включающей в себя архитектуры TDNN и BLSTM.

6. Проведена оценка качества распознавания с использованием разработанных АМ и классификатора фонем, обученных на небольшом объеме данных (около 20 часов). Использование предложенной модели нейросетевой параметризации речевого сигнала позволяет повысить точность распознавания на 2,7% по сравнению с моделью, извлекающей стандартные bottleneck-признаки, а разработанные АМ и архитектура нейросети для распознавания фонем показывают результат распознавания на 4,1% лучше, чем ASR от Google, которая обучалась на речевых базах объемом в десятки тысяч часов.

7. На основе предложенных методов и моделей разработана ASR-система, которая обучалась на речевом корпусе объемом около 7,2 Гб. Авторская система по качеству превосходит решение компании ЦРТ на 10,42%, уступая Google на 6,44%. Разработанная ASR обладает достаточной точностью и превосходит ASR Google и ЦРТ по скорости распознавания более, чем в 7 раз.

8. Модернизация алгоритмов системного анализа, обработки и распознавания речевой информации, а также автоматической обработки текстовых данных позволила повысить эффективность дикторонезависимой системы автоматического распознавания слитной русской речи, работающей с быстродействием и точностью, достаточными для практических задач, и требующей для своего обучения объем данных более, чем в 500 раз меньший, чем существующие аналоги.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

В рецензируемых научных изданиях ВАК ДНР:

1. Пикалёв, Я.С. Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов / Я.С. Пикалёв // Проблемы искусственного интеллекта. – Донецк: ГУ ИПИИ. – 2020. – № 4(19). – С. 45-68.

2. Пикалёв, Я.С. Адаптация нейросетевой модели ALBERT для задачи языкового моделирования / Я.С. Пикалёв, Т.В. Ермоленко // Проблемы искусственного интеллекта. – Донецк: ГУ ИПИИ. – 2020. – № 3(18). – С. 111-122.

3. Пикалёв, Я.С. Система автоматической генерации транскрипций русскоязычных слов-исключений на основе глубокого обучения / Я.С. Пикалёв, Т.В. Ермоленко // Проблемы искусственного интеллекта. – Донецк: ГУ ИПИИ. – 2019. – № 4(15). – С. 35-51.

4. Пикалёв, Я.С. Разработка автоматической системы трансформации английских вставок в русских текстах с применением глубокого обучения / Я.С. Пикалёв // Проблемы искусственного интеллекта. – Донецк: ГУ ИПИИ. – 2019. – № 2(13). – С. 74-86.

В рецензируемых изданиях ВАК РФ:

5. Пикалёв, Я.С. Технология повышения робастности акустической модели в задаче распознавания речи / Я.С. Пикалёв, Т.В. Ермоленко // Известия ЮФУ. Технические науки. – Ростов-на-Дону: ЮФУ. – 2019. – № 7 (209). – С. 45-57.

В изданиях, включенных в систему Российского индекса научного цитирования:

6. Пикалёв, Я. С. Разработка системы автоматического распознавания слитной русскоязычной речи на основе дискриминативного обучения // Информатика и кибернетика. – Донецк: ДонНТУ. – 2018. – №3(13). – С. 61-68.

7. Пикалёв Я. С. Применение систем синтеза речи // Электронные информационные системы. – Москва: НТЦ ЭЛИНС. – 2016. – № 3 (10). – С. 51-56.

По материалам научных конференций:

8. Пикалёв, Я. С. Разработка метода автоматического определения диктора на основе искусственных нейронных сетей для задачи формирования речевой базы // Искусственный интеллект: теоретические аспекты, практическое применение: материалы Донецкого международного научного круглого стола. − Донецк: ГУ ИПИИ, 2020. – С.150-157.

9. Пикалёв, Я. С. Применение аугментации для задачи автоматического распознавания речи / Я.С. Пикалёв, Т.В. Ермоленко // Материалы IV конференции Донецкие чтения 2019: образование, наука, инновации, культура и вызовы современности. Т.1. Ч.2. – 2019. – С. 259-261.

10. Пикалёв, Я. С. Разработка автоматической системы трансформации английских вставок в русских текстах с применением глубокого обучения / Я. С. Пикалёв, Т.В. Ермоленко // Материалы X Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование» (ИУСМКМ-2019). – Донецк, ДОННТУ, 2019. – С. 102-106.

11. Пикалёв, Я.С. Повышение робастности в системе распознавания слов русской слитной речи / Я.С. Пикалёв, Т.В. Ермоленко // XII мультиконференция по проблемам управления (МКПУ-2019): Материалы XII мультиконференции (Дивноморское, Геленджик, 23-28 сентября 2019 г.): в 4 т. / Южный федеральный университет – Ростов-на-Дону; Таганрог: Издательство Южного федерального университета, 2019. – С. 122-125.

12. Пикалёв, Я.С. Глубинное обучение в задаче автоматического распознавания речи // Интеллектуальные технологии и проблемы математического моделирования: Материалы Всерос. науч. конф. (Дивноморское, 24-26 сентября 2018 г.) / под общ. ред. Б.В. Соболя; Донской гос. техн. ун-т. – Ростов-на-Дону: ДГТУ, 2018. – С. 16-17.

13. Пикалёв, Я.С. Разработка синтаксического анализатора русского языка на основе глубоких нейронных сетей // Донецкие чтения 2018: образование, наука, инновации, культура и вызовы современности: Материалы III Международной научной конференции (Донецк, 25 октября 2018 г.). – Т.1. Ч.2. – Донецк: Изд-во ДонНУ, 2018. – С. 240-242.

14. Пикалёв, Я.С. О системах проверки правописания русского языка / Я.С. Пикалёв, А.С. Вовнянко // Донецкие чтения 2018: образование, наука, инновации, культура и вызовы современности: Материалы III Международной научной конференции (Донецк, 25 октября 2018 г.). – Т.1. Ч.2. – Донецк: Изд-во ДонНУ, 2018. – С. 243-247.

15. Пикалёв, Я.С. Классификация текстовых документов при помощи иерархических нейросетей со свёрточным слоем // Восьмая международная конференция по когнитивной науке: Тезисы докладов. (Светлогорск, 18-21 октября 2018 г.) – М.: Изд-во «Институт психологии РАН», 2018. – С. 810-813.

16. Пикалёв, Я.С. Исследование программного комплекса распознавания речи Kaldi ASR / Я.С. Пикалёв, В.Ю. Шелепов //Донецкие чтения 2017: Русский мир как цивилизационная основа научно-образовательного и культурного развития Донбасса: Материалы Международной научной конференции студентов и молодых ученых (Донецк, 17-20 октября 2017 г.). – Т.1. Ч.2. – Донецк: Изд-во ДонНУ, 2017. – С. 232-234.

17. Пикалев, Я.С. Применение систем синтеза речи // Материалы VII Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование» (ИУСМКМ-2016). – Донецк, ДОННТУ, 2016. – С. 136-142.

Участие соискателя в совместных публикациях состоит в следующем: в [2] автором разработана и программно реализована языковая модель на базе модели ALBERT; в [3, 10] – разработаны архитектуры нейросетей и программно реализованы транскрипторы на их основе, а также выполнены практические эксперименты; в[5, 11] – разработаны и программно реализованы робастные акустические модели, проведены численные исследования; в [9] – предложена и программно реализована техника аугментации обучающей речевой выборки; в [14] – разработана и программно реализована нейросетевая модель для проверки правописания, выполнены практические эксперименты; в [16] – программно реализована ASR-система с использованием библиотеки Kaldi и проведены численные исследования эффективности ее работы.

АННОТАЦИЯ

Пикалёв Я. С. Совершенствование методов и программных средств распознавания слитной русской речи. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.01 – Системный анализ, управление и обработка

информации (по отраслям) (технические науки). – ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», Донецк, 2021 г.

Диссертация посвящена совершенствованию моделей и методов обработки и распознавания речевой информации, позволяющих учитывать особенности русского языка, адаптирующихся под любую предметную область, диктора и акустическую обстановку.

Исследованы модели, методы, алгоритмы извлечения акустических признаков, акустического и языкового моделирования. В результате проведенного анализа разработаны:

– методы автоматического формирования словаря транскрипций простых слов, слов-исключений и практической транслитерации;

– методы получения робастных акустических признаков и акустические модели на основе глубоких нейронных сетей;

– нейросетевая модель для классификации фонем.

Создана система, по точности распознавания превосходящая решение компании ЦРТ на 10,42%, уступающая Google на 6,44%.

Модернизация алгоритмов системного анализа, обработки и распознавания речевой информации, автоматической обработки текстовых данных позволила повысить эффективность дикторонезависимой системы автоматического распознавания слитной русской речи, работающей с быстродействием и точностью, достаточными для практических задач, и требующей для своего обучения объем данных более, чем в 500 раз меньший, чем существующие аналоги.

Ключевые слова: система распознавания речи, нейронные сети, обработка и распознавание речевой информации, акустическая и языковая модель, качество распознавания.

ANNOTATION

Pikaliov Y. S. Improvement of methods and software for continuous Russian speech recognition. – As a manuscript.

Thesis for the degree of candidate of technical sciences in specialty 05.13.01 – System analysis, management and information processing (by industry) (technical sciences). – STATE HIGHER EDUCATION ESTABLISHMENT “DONETSK NATIONAL TECHNICAL UNIVERSITY”, Donetsk, 2021.

The thesis is devoted to the improvement of models and methods of processing and recognition of speech information, allowing to take into account the peculiarities of the Russian language, adapting to any subject area, speaker and acoustic environment.

Models, methods, algorithms for extracting acoustic features, acoustic and language modeling have been investigated. As a result of the analysis, the following were developed:

– methods of automatic formation of a dictionary of transcriptions of simple words, words-exceptions and practical transliteration;

– methods for obtaining robust acoustic features and acoustic models based on deep neural networks;

– neural network model for classification of phonemes.

A system has been created that surpasses the solution of the MDG company in recognition accuracy by 10.42%, yielding to Google by 6.44%.

Modernization of algorithms for system analysis, processing and recognition of speech information, automatic processing of textual data has made it possible to increase the efficiency of the speaker-independent system for automatic recognition of continuous Russian speech, which operates with speed and accuracy sufficient for practical tasks, and requires more than 500 times the amount of data for its training. smaller than existing counterparts.

Key words: speech recognition system, neural networks, processing and recognition of speech information, acoustic and language model, recognition quality.