На правах рукописи

Меденников Иван Павлович

МЕТОДЫ, АЛГОРИТМЫ И ПРОГРАММНЫЕ СРЕДСТВА РАСПОЗНАВАНИЯ РУССКОЙ ТЕЛЕФОННОЙ СПОНТАННОЙ РЕЧИ

Специальность 05.13.11 —
«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург — 2016

Работа выполнена на кафедре теории управления федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет»

Научный руководитель: **Жабко Алексей Петрович**
доктор физико-математических наук, профессор, профессор с возложенными обязанностями заведующего кафедрой теории управления федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет», заслуженный работник Высшей школы Российской Федерации

Официальные оппоненты: **Левин Евгений Калманович**
доктор технических наук, доцент, профессор кафедры радиотехники и радиосистем федерального государственного бюджетного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»

Клионский Дмитрий Михайлович
кандидат технических наук, доцент кафедры математического обеспечения и применения ЭВМ, заместитель декана факультета компьютерных технологий и информатики по международной деятельности федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В. И. Ульянова (Ленина)»

Ведущая организация: Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)

Защита состоится «29» сентября 2016 г. в 15 часов 30 минут на заседании диссертационного совета Д 212.232.51 на базе Санкт-Петербургского государственного университета по адресу: 198504, Санкт-Петербург, Старый Петергоф, Университетский пр., 28, математико-механический факультет, ауд. 405.

С диссертацией можно ознакомиться в Научной библиотеке им. М. Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб., 7–9 и на сайте https://disser.spbu.ru.

Автореферат разослан «___» _____ 2016 года.

Ученый секретарь диссертационного совета
Д 212.232.51, д. ф.-м. н., профессор  Демьянович Юрий Казимирович

Общая характеристика работы

Актуальность темы исследования. Развитие науки и техники неразрывно связано с эволюцией средств взаимодействия человека и машины. В современном мире все большую популярность приобретает речевой интерфейс человеко-машинного взаимодействия. Это обусловлено тем, что именно речь является наиболее естественным для человека средством коммуникации. Важнейшей составляющей речевого человеко-машинного интерфейса являются системы автоматического распознавания речи.

Одной из наиболее сложных задач в области автоматического распознавания речи является задача распознавания разговорной спонтанной речи — стиля речи, который характеризуется отсутствием заранее подготовленной формы и содержания устного сообщения и непосредственным участием говорящих. Сложность задачи вызвана следующими особенностями разговорной спонтанной речи: значительная междикторская вариативность, вариативность темпа речи и манеры произнесения, наличие акцентной и эмоциональной речи, большое количество используемых словоформ. Задача дополнительно осложняется наличием хезитаций — речевых колебаний, связанных со спонтанностью речи, к которым относятся паузы, нелексические вставные звуки, «слова-паразиты», коррекции предложения, замены слов, повторы, заикания, незавершенные предложения. В ситуации речевой коммуникации именно спонтанная речь является первичной, поэтому задача ее распознавания крайне актуальна.

Системы распознавания телефонной спонтанной речи являются крайне востребованными, например, в задачах контроля качества обслуживания в контакт-центрах и анализа тематик больших архивов телефонных переговоров. Однако при использовании телефонного канала имеются различные особенности, ухудшающие качество работы систем распознавания речи. К ним относятся ограничение полосы пропускания диапазоном частот 0–4000 Гц, наличие аддитивных и нелинейных канальных искажений, а также потеря информации в результате кодирования речевого сигнала. Эти особенности дополнительно осложняют задачу распознавания телефонной спонтанной речи.

Актуальность темы исследования подтверждается большим количеством посвященных ей докладов на международных конференциях, таких как Interspeech, ICASSP, SPECOM, ASRU, TSD, а также повсеместным внедрением систем автоматического распознавания спонтанной речи.

Степень разработанности темы исследования. Для исследований по распознаванию английской спонтанной речи используются корпуса телефонных разговоров на английском языке Switchboard-1 (300 часов), корпус Фишера (2000 часов) и другие. Исследованиям, проведенным на этих базах, посвящено большое количество работ ученых из IBM (Brian Kingsbury, George Saon и др.), Microsoft (Li Deng, Dong Yu, Frank Seide и др.), Университета Торонто (George E. Dahl и др.), Университета Джона Хопкинса (Daniel Povey и др.),

Google (Andrew Senior, Tara Sainath и др.) и других исследовательских коллективов. Построенные в этих работах системы распознавания обладают высоким качеством, которое позволяет применять их в коммерческих продуктах. Лучшие на сегодняшний день системы распознавания английской телефонной спонтанной речи обеспечивают уровень ошибки распознавания около 15%.

Распознаванию слитной и спонтанной русской речи посвящены работы исследователей из Санкт-Петербургского института информатики и автоматизации Российской академии наук (Андрей Ронжин, Алексей Карпов, Ирина Кипяткова и др.), компании ООО «ЦРТ» (Михаил Хитров, Кирилл Левин, Максим Кореневский, Юрий Хохлов, Марина Татарникова и др.), Университета ИТМО (Иван Тампель и др.), лаборатории LIMSI (Франция) (Lori Lamel и др.), а также исследовательских коллективов компаний Яндекс, Google, Phonexia (Чехия) и других.

В 2014 году Фондом Перспективных Исследований (ФПИ) был организован конкурс-семинар по распознаванию речи, целью которого являлось определение российских фирм-разработчиков, обладающих в настоящее время наиболее эффективными аппаратно-программными решениями по преобразованию речи в текст. Одна из его секций была посвящена дикторонезависимому распознаванию русской телефонной спонтанной речи. В конкурсе приняли участие следующие компании: ООО «ЦРТ» (победитель), ФГУП «НИИ «Квант», ООО «Стэл-КС», ЗАО «НТЦ «Поиск-ИТ». Стоит отметить, что даже система-победитель конкурса ФПИ демонстрирует недостаточно высокую точность распознавания русской телефонной спонтанной речи — по результатам распознавания, полученным с ее помощью, во многих случаях не удается восстановить смысл сказанного. Таким образом, на настоящий момент не существует систем распознавания русской спонтанной речи, сопоставимых по качеству с вышеупомянутыми системами для английского языка.

Можно выделить несколько причин недостаточной эффективности существующих систем распознавания русской телефонной спонтанной речи. Во-первых, в открытом доступе отсутствуют обучающие корпуса записей русской телефонной спонтанной речи и общепринятые базы для оценки качества систем распознавания русской спонтанной речи. Во-вторых, русский язык, относящийся к флективным языкам, имеет существенно большее число словоформ, по сравнению с аналитическими языками. Вышеупомянутые системы распознавания английской спонтанной речи оперируют словарями объемом несколько десятков тысяч слов, в то время как для эффективной работы системы распознавания русской разговорной речи необходим словарь, содержащий сотни тысяч слов. В-третьих, задачу усложняют фонетические особенности русской спонтанной речи, а именно вялая артикуляция, явления ассимиляции (объединения звуков) и редуцирования (сокращения длительности звуков). Эффективная система распознавания русской спонтанной речи должна быть устойчивой к акустической вариативности речевого сигнала, вызванной этими фонетическими особенностями.

Учитывая вышесказанное, можно сделать вывод о необходимости разработки методов, алгоритмов и программных средств, обеспечивающих повышение точности распознавания русской телефонной спонтанной речи.

Целью данной работы является разработка методов, алгоритмов и программных средств, позволяющих повысить точность распознавания русской телефонной спонтанной речи, и их реализация в системе, работающей с быстродействием, достаточным для применения в практических задачах. Для достижения поставленной цели были сформулированы и решены следующие основные **задачи**.

1. Анализ современных методов распознавания спонтанной речи.
2. Разработка методов, алгоритмов и программных средств распознавания русской телефонной спонтанной речи.
3. Построение языковой модели, словаря транскрипций и акустической модели, входящих в состав системы распознавания русской телефонной спонтанной речи.
4. Оценка качества работы разработанной системы распознавания русской телефонной спонтанной речи, а также сравнение с российскими и зарубежными системами.

Объект исследования. Системы автоматического распознавания речи.

Предмет исследования. Методы, алгоритмы и программные средства автоматического распознавания русской телефонной спонтанной речи.

Используется широко распространенная в прикладных научных исследованиях **методология**: формулирование целей и задач, анализ состояния исследований и существующей литературы, разработка алгоритмических и программных решений, экспериментальная оценка эффективности разработанных решений, апробация и анализ результатов. Особое внимание следует уделить методологии проведения экспериментальной части исследования — она проводилась исключительно на естественном речевом материале, при этом тестовые выборки ни по произнесениям, ни по составу дикторов не пересекались с обучающими данными. В качестве **методов исследования** используются методы цифровой обработки сигналов, теории вероятностей и математической статистики, машинного обучения, прикладной лингвистики, а также методы разработки программного обеспечения.

Научная новизна.

1. Разработан метод построения информативных признаков, извлекаемых из глубокой нейронной сети с узким горлом, отличающийся применением адаптации к диктору и акустическим условиям и позволяющий улучшить качество акустических моделей для спонтанной речи.
2. Разработан двухэтапный алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей, отличающийся учетом количества неречевых примеров в обучающей выборке и обеспечивающий повышение точности распознавания спонтанной речи.

3. Разработан метод построения системы распознавания русской телефонной спонтанной речи, включающий в себя обучение языковых моделей, формирование словаря транскрипций и обучение акустических моделей с использованием разработанных метода и алгоритма.
4. Реализованы программные средства, входящие в состав системы распознавания русской телефонной спонтанной речи и позволяющие использовать акустические модели, построенные с помощью представленных в диссертации методов и алгоритмов.

Теоретическая и практическая значимость работы. Теоретическая значимость данной работы заключается в улучшении существующих и разработке новых алгоритмов обучения акустических моделей на основе глубоких нейронных сетей для задачи распознавания речи, а также в разработке и экспериментальном исследовании нового метода извлечения информативных признаков, превосходящего использовавшиеся ранее.

Практическая значимость диссертационного исследования заключается в использовании разработанных алгоритмических и программных средств при создании системы распознавания русской телефонной спонтанной речи, демонстрирующей достаточно высокие качество распознавания и быстродействие для применения в таких практических задачах, как автоматическая отекстовка архивов фонограмм, поиск ключевых слов в потоке слитной речи, кластеризация записей по тематикам. Основные результаты, полученные в диссертации, внедрены:

1. В состав ряда коммерческих продуктов компании ООО «ЦРТ»: АПК «Трал», ПО «VoiceNavigator», ПО «VoiceNavigator Web», ПО «Незабудка II».
2. В компании ООО «ЦРТ» при выполнении научно-исследовательских и опытно-конструкторских работ по теме «Разработка аппаратно-программного комплекса автоматической подготовки скрытых субтитров в реальном масштабе времени для внедрения на общероссийских обязательных общедоступных телеканалах в пределах утвержденных лимитов бюджетных обязательств» в рамках выполнения обязательств по Государственному контракту от 7 декабря 2012 г. № 0173100007512000034_144316, а также при выполнении составной части проекта по теме «Модернизация речевого сервера для использования в макете перспективной системы транскрибирования речи. Разработка систем тематического рубрицирования и дообучения к источнику речи» шифр «Лангет-Ц».
3. В компании ООО «ЦРТ-инновации» при проведении прикладных научных исследований по теме «Разработка технологии преобразования русской речи в транскрипционное представление с метаданными для автоматического распознавания речевых команд в робототехнике и промышленности» в рамках Соглашения с Министерством образования и науки РФ № 14.579.21.0057 от 23.09.2014 (ID проекта RFMEFI57914X0057), а так-

же прикладных научных исследований по теме «Разработка методов лингвистического и семантического анализа для интеллектуальной обработки текстов, полученных в результате автоматического распознавания звучащей спонтанной русской речи» в рамках Соглашения с Министерством образования и науки РФ № 14.579.21.0008 от 5 июня 2014 г. (ID проекта RFMEFI57914X0008).

Основные положения, выносимые на защиту:

1. Метод построения информативных признаков, извлекаемых из адаптированной к диктору и акустическим условиям глубокой нейронной сети с узким горлом.
2. Двухэтапный алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей.
3. Метод построения системы распознавания русской телефонной спонтанной речи.
4. Программные средства, входящие в состав системы распознавания русской телефонной спонтанной речи.

Степень достоверности и апробация результатов. Достоверность и обоснованность результатов исследования обеспечивается корректным обоснованием постановок задач, точной формулировкой критериев, анализом состояния исследований в данной области, проведением большого количества экспериментов, а также успешным внедрением на практике. Результаты диссертации докладывались и обсуждались на следующих научно-методических конференциях: «15th Annual Conference of the International Speech Communication Association» (Сингапур, 2014), «16th International Conference on Speech and Computer» (Нови Сад, Сербия, 2014), «17th International Conference on Speech and Computer» (Афины, Греция, 2015), «XLV научная и учебно-методическая конференция Университета ИТМО» (Санкт-Петербург, Россия, 2016).

Личный вклад автора. Соискателем лично решены задачи диссертации. Разработаны методы и алгоритмы распознавания спонтанной речи, проведена экспериментальная оценка эффективности разработанных методов и алгоритмов. Разработаны программные средства, входящие в состав системы распознавания русской телефонной спонтанной речи.

Публикация результатов. По теме диссертации опубликовано семь печатных работ. Статьи [1], [2] опубликованы в журналах из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук. Статьи [3–7] опубликованы в изданиях, индексируемых в международных реферативных базах Scopus или Web of Science. В статье [3] соискателю принадлежит построение языковых моделей, проведение экспериментов по подбору параметров декодера и настройка быстродействия системы. В статье [4] соискателю принадлежит построение языковых моделей, построение акустической модели для распознавания рус-

ской телефонной спонтанной речи, а также получение базовых результатов по скорости и точности распознавания. В статье [5] соискателю принадлежит построение языковой модели и настройка параметров декодера. В статье [6] соискателю принадлежит разработка метода построения признаков, извлекаемых из адаптированной к диктору и акустическим условиям глубокой нейронной сети, построение языковой модели, построение акустических моделей, проведение экспериментов по оценке эффективности разработанного метода. В статье [7] соискателю принадлежит построение акустической модели для автоматического распознавания казахской и русской речи. Остальные результаты в статьях [3–7] принадлежат соавторам.

Объем и структура работы. Диссертация состоит из введения, четырех глав и заключения. Полный объём диссертации составляет 148 страниц с 18 рисунками и 32 таблицами. Список литературы содержит 146 наименований.

Основное содержание работы

Во введении обоснована актуальность исследования, проводимого в рамках данной диссертационной работы, сформулированы цели и задачи исследования, показаны научная новизна и практическая значимость работы, представлены положения, выносимые на защиту.

В первой главе приведен обзор современных методов автоматического распознавания речи. Рассмотрена типичная структура современной системы автоматического распознавания речи, в состав которой входят модуль обработки сигнала и извлечения признаков, акустическая модель, языковая модель, словарь транскрипций и декодер. Модуль обработки сигнала и извлечения признаков принимает на вход звуковой сигнал, осуществляет шумоочистку и извлекает векторы информативных признаков, которые в дальнейшем используются при акустическом моделировании. Акустическая модель описывает плотность распределения вероятностей акустических классов (например, фонем) на заданном участке речевого сигнала. Языковая модель описывает вероятность появления слова в контексте других слов. Словарь транскрипций устанавливает связь между последовательностями акустических классов, описываемых акустической моделью, и словами, описываемыми языковой моделью. Наконец, декодер анализирует вероятности, генерируемые акустической и языковой моделями, и выдает в качестве результата распознавания последовательность слов $\hat{w}$, определяемую как

$$\hat{w} = \arg\max_{w} \mathsf{P}(w|\mathbf{x}) = \arg\max_{w} \frac{\mathsf{P}(\mathbf{x}|w)\,\mathsf{P}(w)}{\mathsf{P}(\mathbf{x})} = \arg\max_{w} \mathsf{P}(\mathbf{x}|w)\,\mathsf{P}(w). \qquad (1)$$

Здесь максимум берется по всем возможным цепочкам слов w, $\mathbf{x}$ представляет собой набор векторов признаков распознаваемого сигнала, $\mathsf{P}(w)$ — генериру-

емая языковой моделью вероятность цепочки слов w, а $\mathsf{P}(\mathbf{x}|w)$ — вероятность, генерируемая акустической моделью.

Далее в тексте главы проанализированы особенности разговорной русской речи и проведен анализ состояния исследований в области распознавания диктовочной и спонтанной речи на русском языке. Сделан вывод, что на настоящий момент существуют системы, успешно решающие задачу распознавания диктовочной русской речи, однако не разработано систем, эффективно распознающих спонтанную русскую речь.

Вторая глава посвящена исследованию и разработке методов и алгоритмов построения информативных признаков при помощи глубоких нейронных сетей (Deep Neural Networks, DNN), а также акустических моделей на основе DNN и скрытых марковских моделей (Hidden Markov Models, HMM).

В начале главы приводится интерпретация DNN как составной модели, совмещающей каскад нелинейных преобразований входных признаков и классификатор. Приведены результаты исследований, показывающих, что нелинейные преобразования признаков, осуществляющиеся на скрытых слоях DNN, обеспечивают устойчивость по отношению к малым возмущениям входного сигнала. Дано описание DNN с узким горлом, позволяющих извлекать признаки, обладающие устойчивостью по отношению к акустической вариативности речевого сигнала.

Далее представлен разработанный автором метод построения высокоуровневых информативных признаков, идея которого заключается в использовании адаптированной DNN для извлечения признаков. Основой для этой идеи послужил сделанный в главе вывод о том, что чем лучше точность распознавания, которая обеспечивается DNN с узким горлом, тем лучшую точность распознавания будет обеспечивать система, построенная на основе признаков, извлеченных из этой DNN. Анализ алгоритмов адаптации DNN-HMM акустических моделей, проведенный в первой главе, показал, что адаптация DNN к диктору и акустической обстановке с использованием i-векторов — малоразмерных векторов, кодирующих отличие плотности распределения вероятностей акустических признаков, оцененной по фонограмме, от эталонной — значительно повышает точность распознавания речи за счет предоставления DNN дополнительной информации о фонограмме. Таким образом, в основе разработанного метода лежит предположение, что использование признаков, извлекаемых из DNN с узким горлом, адаптированной при помощи i-векторов, позволит повысить устойчивость по отношению к акустической вариативности и обеспечить лучшую способность к разделению акустических классов.

Основные этапы представленного в диссертации алгоритма построения признаков, согласно разработанному методу, таковы:

1. Обучение неадаптированной DNN по критерию минимизации взаимной энтропии (рисунок 1а).
2. Расширение входного слоя обученной DNN с инициализацией соответствующих коэффициентов матрицы весов нулевыми значениями.

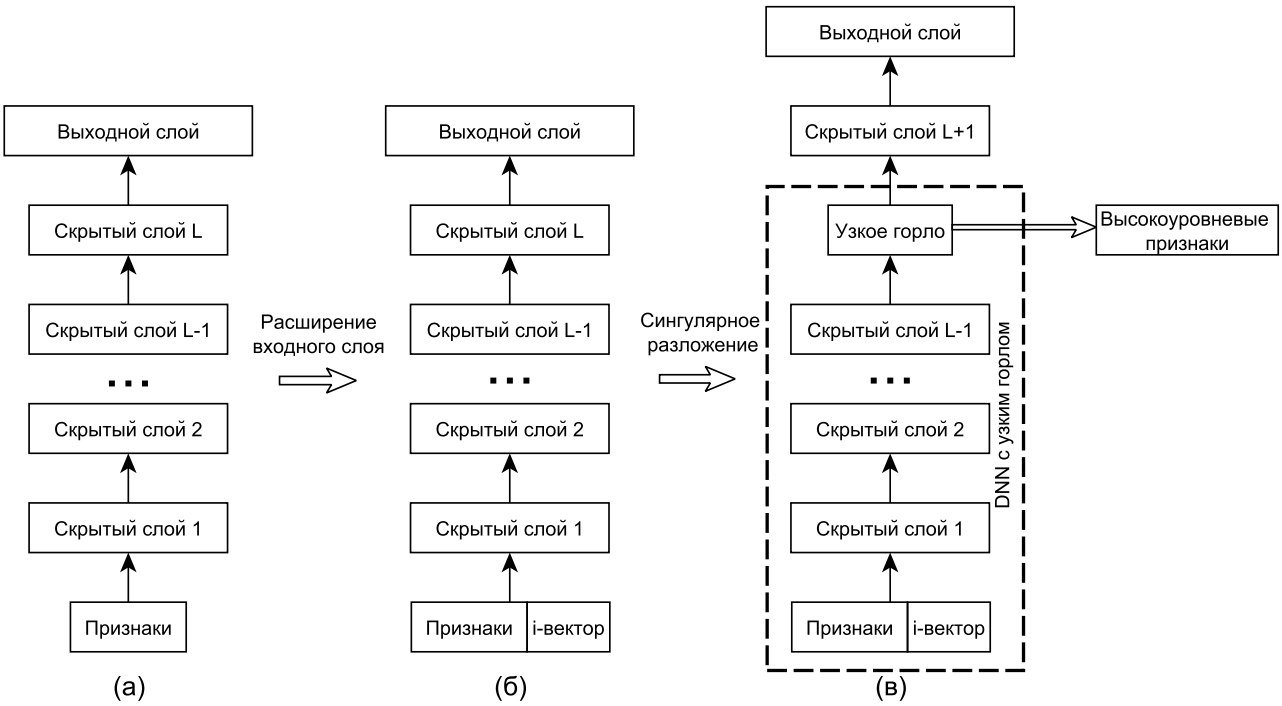

Рисунок 1 — Основные этапы обучения глубокой нейронной сети с узким горлом, адаптированной при помощи i-векторов.

3. Дообучение DNN с расширенным входным слоем по признакам, к которым на каждом кадре добавлен i-вектор, соответствующий данному участку фонограммы (рисунок 1б). При этом используется меньшая скорость обучения, а к целевой функции добавлено слагаемое $R(\mathbf{W})$, штрафующее отклонение весов $\mathbf{W}^l$ обучаемой модели от значений весов $\bar{\mathbf{W}}^l$ исходной модели, определяемое по формуле

$$R(\mathbf{W}) = \lambda \sum_{l=1}^{L+1} \sum_{i=1}^{N_l} \sum_{j=1}^{N_{l-1}} (\mathbf{W}_{ij}^l - \bar{\mathbf{W}}_{ij}^l)^2, \tag{2}$$

где λ — величина штрафа.

4. Разбиение слоя l глубокой нейронной сети (например, последнего скрытого слоя) на два слоя следующим образом:

$$\mathbf{v}^l = f(\mathbf{W}^l \mathbf{v}^{l-1} + \mathbf{b}^l) \approx f(\mathbf{W}_{out}^l (\mathbf{W}_{bn}^l \mathbf{v}^{l-1} + \mathbf{0}) + \mathbf{b}^l). \tag{3}$$

Здесь первый слой — малоразмерный слой с линейной функцией активации, матрицей весов $\mathbf{W}_{bn}^l$ и нулевым вектором смещений; второй слой — нелинейный слой с матрицей весов $\mathbf{W}_{out}^l$ и вектором смещений $\mathbf{b}^l$, имеющий размерность исходного разбиваемого слоя. Разбиение осуществляется при помощи сингулярного разложения (Singular Values Decomposition,

SVD) матрицы весов $\mathbf{W}^l$:

$$\mathbf{W}^l \approx \mathbf{W}^l_{out}\mathbf{W}^l_{bn}. \quad (4)$$

Таким образом, исходная DNN с L скрытыми слоями преобразуется в DNN с $(L+1)$ скрытыми слоями с линейным узким слоем l.

5. Дообучение полученной DNN с узким горлом (рисунок 1в) с меньшей скоростью и штрафом на отклонение весов от весов исходной модели, определяемым по формуле 2.
6. Отбрасывание слоев DNN, следующих за узким горлом, и использование полученной DNN с узким горлом для построения высокоуровневых информативных признаков.

Предлагается следующий алгоритм обучения DNN-HMM акустических моделей на основе информативных признаков, построенных согласно описанному выше методу:

1. Обучение трифонной акустической модели на основе моделей гауссовых смесей (Gaussian Mixture Models, GMM) и HMM с использованием построенных признаков.
2. Разметка обучающих данных на связанные состояния трифонов при помощи построенной трифонной GMM-HMM модели.
3. Обучение DNN-HMM модели с использованием построенных признаков, взятых с широким временным контекстом.

Также во второй главе представлен двухэтапный алгоритм инициализации обучения акустических моделей на основе DNN. Основой для него послужило наблюдение, что сегменты, не содержащие речи, составляют значительную долю в фонограммах, на которых осуществляется обучение акустических моделей. По этой причине при обучении DNN по критерию минимизации взаимной энтропии может возникать ситуация, когда качество классификации неречевых фонем улучшается в ущерб качеству классификации речевых фонем, и, следовательно, в ущерб качеству распознавания речи. Предложенный алгоритм направлен на уменьшение влияния этого эффекта и состоит из двух этапов:

1. Осуществляется предобучение DNN одним из способов: при помощи ограниченных машин Больцмана, автоэнкодеров, или дискриминативного алгоритма предобучения.
2. Полученная на первом этапе предобученная DNN используется для инициализации обучения по критерию минимизации взаимной энтропии на сбалансированной по количеству неречевых примеров обучающей выборке. Балансировка происходит следующим образом: из обучающих примеров, соответствующих неречевым фонемам, случайным образом выбирается некоторая их часть так, чтобы количество примеров для неречевых фонем в обучающей выборке было примерно равным среднему количеству примеров для одной речевой фонемы.

DNN, полученную на втором этапе алгоритма, в дальнейшем предлагается использовать для инициализации обучения по полной обучающей выборке. Это способствует улучшению качества классификации неречевых фонем без большого ущерба для качества классификации речевых фонем, что позволяет повысить точность распознавания речи. Также предложены варианты использования предложенного двухэтапного алгоритма для обучения DNN, адаптированных при помощи i-векторов.

В **третьей главе** представлен метод построения системы распознавания русской телефонной спонтанной речи, включающий в себя обучение языковых моделей, формирование словаря транскрипций и обучение акустических моделей.

Для построения системы распознавания использовался 400-часовой обучающий корпус, состоящий из записей телефонной спонтанной речи на русском языке. Все фонограммы были записаны с частотой дискретизации 8000 Гц, 16 бит на отсчет. Записи характеризовались большой дикторской вариативностью, а также разнообразием акустической обстановки, в которой происходили записываемые диалоги. Для настройки системы и экспериментов использовались четыре базы длительностью 30 минут, 1 час 18 минут, 1 час 43 минуты, 44 минуты соответственно, не пересекающиеся ни по произнесениям, ни по составу дикторов с обучающими данными.

Триграммная языковая модель с модифицированным сглаживанием Кнесера-Нея была построена по текстовым расшифровкам записей обучающего корпуса, дополненным текстовыми данными, собранными из открытых источников: субтитров к фильмам, современных книг и текстов обсуждений с форумов сети Интернет. Перед построением языковой модели тексты подвергались нормализации — автоматической очистке от спецсимволов, опечаток и орфографических ошибок. Построенная языковая модель содержала 214 тыс. униграмм, 4 млн. биграмм и 2,4 млн. триграмм.

Транскрипции, или последовательности фонем, соответствующие слову, были сгенерированы автоматически с использованием инструмента — транскриптора, разработанного в ООО «ЦРТ». Всего для списка из 214 тыс. слов, содержащихся в языковой модели, было сгенерировано 220 тыс. канонических транскрипций, отражающих произнесение слова с точки зрения норм русского языка. Поскольку произношение слов в русской спонтанной речи зачастую значительно отличается от канонического в силу эффектов ассимиляции и редукции звуков, а также других особенностей произношения в разговорной речи, существует необходимость добавления неканонических, или альтернативных, транскрипций в словарь. Для тысячи наиболее частотных в языковой модели слов были вручную добавлены альтернативные транскрипции.

Для учета эффектов коартикуляции (взаимного влияния звуков в слитной речи друг на друга) в словосочетаниях, а также фонетических особенностей русской спонтанной речи, автором был предложен двухпроходный алгоритм распознавания речи, продемонстрировавший потенциал для улучшения точ-

ности распознавания. Однако необходимость второго прохода распознавания значительно (на 30–40%) замедляет работу системы, поэтому двухпроходный алгоритм не был использован в разработанной системе распознавания русской телефонной спонтанной речи.

Первым этапом построения акустических моделей для системы распознавания русской телефонной спонтанной речи было прохождение пути, аналогичного рецепту для английской спонтанной речи из инструмента Kaldi ASR. Далее было проведено обучение DNN-HMM акустических моделей (шесть скрытых слоев по 1024 нейрона), адаптированных при помощи i-векторов. Адаптация DNN-HMM модели при помощи i-векторов позволила достичь 2,1–2,6% абсолютного (абс.) и 4,7–6,6% относительного (отн.) улучшения показателя пословной ошибки распознавания (Word Error Rate, WER), являющегося общепринятым критерием качества работы системы распознавания речи и определяемого по формуле

$$\mathrm{WER} = \frac{S + I + D}{N} \cdot 100\% = \frac{S + I + D}{C + S + D} \cdot 100\%, \tag{5}$$

где N — количество слов в эталонном тексте, C — количество правильно распознанных слов, S, I, D — соответственно число замен, вставок и удалений в результате распознавания. Использование представленного во второй главе двухэтапного алгоритма инициализации обучения позволило добиться дополнительного улучшения: 2,9–4,0% абс. и 6,2–9,1% отн. превосходства по WER над базовой неадаптированной моделью.

Для дальнейшего улучшения достигнутых результатов был использован представленный во второй главе метод извлечения высокоуровневых информативных признаков из адаптированной при помощи i-векторов DNN. Высокоуровневые признаки были построены по алгоритму, представленному во второй главе, и затем использовались для обучения DNN-HMM модели (четыре скрытых слоя по 2048 нейронов). Результаты оценки эффективности построенной модели говорят о превосходстве по WER над адаптированной при помощи i-векторов DNN-HMM моделью на 0,7–2,5% абс. и 2,6–5,3% отн. Это подтверждает высокую эффективность предложенного во второй главе метода построения высокоуровневых информативных признаков в задаче распознавания русской телефонной спонтанной речи.

Далее было проведено экспериментальное исследование, нацеленное на поиск эффективной конфигурации «сырых» признаков для обучения DNN-HMM акустических моделей. По результатам сравнения семи различных конфигураций признаков была найдена лучшая конфигурация, использование которой позволило добиться 1,5–4,8% абс. и 4,5–9,4% отн. улучшения показателя WER по сравнению с конфигурацией признаков, использованной ранее. Эта конфигурация признаков была использована для построения новых высокоуровневых признаков, на которых затем была обучена финальная DNN-HMM

акустическая модель. Улучшение WER, достигнутое за счет перехода на новую конфигурацию «сырых» признаков, составило 3,4–3,9% абс. и 7,2–13,7% отн.

Для демонстрации суммарной эффективности использованных техник, а именно выбора конфигурации «сырых» признаков, адаптации DNN при помощи i-векторов, использования признаков, извлекаемых из адаптированной DNN с узким горлом, а также применения двухэтапного алгоритма инициализации обучения DNN-HMM моделей, в таблице 1 приведены результаты сравнения финальной акустической модели с базовой DNN-HMM моделью — лучшей из моделей, построенных без использования вышеперечисленных техник.

Таблица 1 — Оценка суммарной эффективности техник, использованных при построении финальной акустической модели

Акустическая модель	WER, %			
	база 1	база 2	база 3	база 4
базовая	28,5	45,3	49,3	46,9
финальная	22,0	37,3	41,4	38,1

В четвертой главе дано описание программных средств разработанной системы распознавания русской телефонной спонтанной речи, приведены результаты сравнения с существующими системами распознавания слитной русской речи по точности распознавания, а также проведена оценка быстродействия разработанной системы.

Разработанная система распознавания русской телефонной спонтанной речи состоит из двух основных подсистем: подсистемы обучения (отвечает за создание акустических и языковых моделей, а также словаря транскрипций) и подсистемы распознавания речи (осуществляет автоматическое преобразование речи из входных wav-файлов в текст, используя при этом результаты работы подсистемы обучения).

Схема подсистемы обучения представлена на рисунке 2. Эта подсистема отвечает за построение следующих составных частей системы распознавания речи: DNN с узким горлом, акустическая модель, языковая модель, словарь транскрипций. Программные средства, входящие в состав подсистемы обучения, были реализованы автором с использованием языков программирования C++, Perl, Python, Bash, и позволили реализовать методы и алгоритмы, представленные в диссертации.

Схема подсистемы распознавания речи представлена на рисунке 3. Данная подсистема принимает на вход фонограмму с частотой дискретизации 8000 Гц, 16 бит на отсчет. Первым этапом обработки входного сигнала является выделение речевых сегментов при помощи детектора активности диктора. По выделенным речевым сегментам происходит вычисление i-векторов, а также вычисление признаков. Построенный на каждом кадре вектор признаков и i-вектор, соответствующий участку фонограммы, которому принадлежит

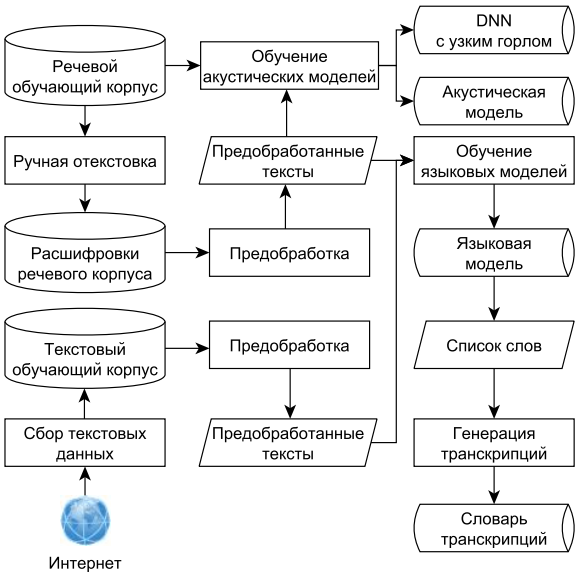

Рисунок 2 — Схема подсистемы обучения

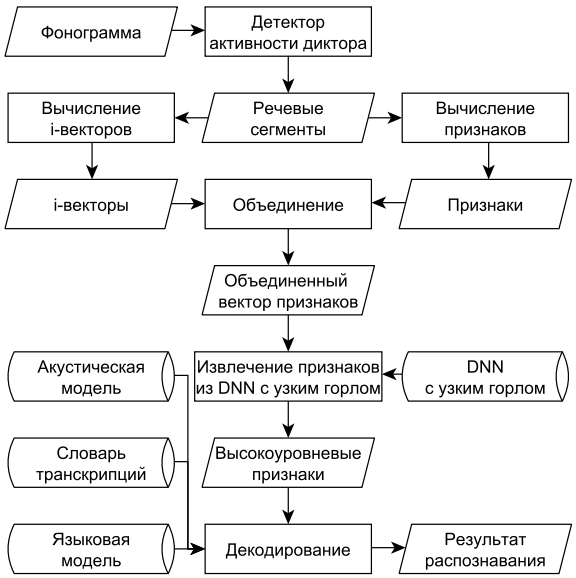

Рисунок 3 — Схема подсистемы распознавания речи

рассматриваемый кадр, объединяются в единый вектор признаков. По объединенным векторам признаков вычисляются векторы высокоуровневых признаков при помощи DNN с узким горлом. Далее осуществляется декодирование, или поиск наиболее правдоподобной последовательности слов, соответствующей последовательности векторов высокоуровневых признаков для данной фонограммы. При декодировании используются акустическая модель, языковая модель и словарь транскрипций, полученные в результате работы подсистемы обучения. Выдаваемая в процессе декодирования последовательность слов (результат распознавания) записывается в выходной текстовый файл.

В основе этой подсистемы лежит программное средство ASR SDK, разработанное в ООО «ЦРТ» при участии автора. Программное средство реализовано на языке программирования C++ с использованием объектно-ориентированного подхода. Поддерживаются операционные системы Linux CentOS 5.1 и MS Windows XP/7/8 с архитектурой процессора x86 и x64. Программное средство поддерживает ускорение выполнения вычислительных операций с использованием вычислений общего назначения на графических процессорах (General-purpose computing for graphics processing units, GPGPU) при помощи технологии Nvidia CUDA.

В конце главы проводится оценка эффективности разработанной системы. Представлены результаты сравнения по WER с различными системами слитного распознавания на русском языке, а именно с системой-победителем конкурса ФПИ, а также с двумя локальными коммерческими системами распознавания от российского и зарубежного производителей, и двумя системами облачного распознавания на удаленном сервере от российского и зарубежного производителей. По результатам сравнения разработанная система продемон-

стрировала WER на уровне 21,9–39,5% на различных тестовых базах, превзойдя лучшую из участвовавших в сравнении систем на 18,1–21,0% абс. и 34,7–45,3% отн.

Проведена также оценка быстродействия разработанной системы на ЭВМ с процессором Intel Core i5 4570 (таблица 2). В качестве критерия быстродействия использовался real-time factor (RTF) — величина, определяемая как отношение времени, затраченного на распознавание фонограмм, к суммарной длительности распознаваемых фонограмм. Значение RTF менее единицы озна-

Таблица 2 — Оценка быстродействия разработанной системы

Число потоков	RTF (без GPGPU)	RTF (с GPGPU)
1	0,51	0,28
4	0,18	0,10

чает, что распознавание осуществляется быстрее, чем воспроизведение той же записи, что является требованием к скорости работы системы во многих практических задачах. Представлен механизм регулирования быстродействия системы за счет изменения параметров декодера, позволяющий добиться требуемого пользователю соотношения «скорость-качество» и тем самым удовлетворить требованиям по быстродействию, диктуемым реальными приложениями.

Заключение

Итоги выполненного исследования. В диссертации получены следующие основные результаты:

1. Разработан метод построения информативных признаков, извлекаемых из адаптированной к диктору и акустическим условиям глубокой нейронной сети с узким горлом.
2. Разработан двухэтапный алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей, предназначенный для уменьшения влияния сегментов, не содержащих речь, на обучение акустической модели.
3. Разработан метод построения системы распознавания русской телефонной спонтанной речи, включающий в себя обучение языковых моделей, формирование словаря транскрипций и обучение акустических моделей с использованием разработанных метода и алгоритма.
4. Реализованы программные средства, входящие в состав системы распознавания русской телефонной спонтанной речи и позволяющие использовать акустические модели, обученные с использованием представленных в диссертации методов и алгоритмов.

Представленные в диссертации методы, алгоритмы и программные средства были реализованы в системе распознавания русской телефонной спонтанной

речи, обеспечивающей значительно более высокую точность распознавания по сравнению с существующими системами, при этом удовлетворяя диктуемым реальными приложениями требованиям по быстродействию. В частности, разработанная система продемонстрировала на 18,1–21,0% абсолютных и 34,7–45,3% относительных меньшую пословную ошибку распознавания, чем система-победитель конкурса ФПИ.

Рекомендации по применению результатов работы:

1. При использовании разработанной системы распознавания русской телефонной спонтанной речи в практических задачах следует использовать предусмотренный в ней механизм регулирования быстродействия, чтобы обеспечить необходимую скорость работы.
2. Разработанные в диссертации метод построения информативных признаков, извлекаемых из глубокой нейронной сети с узким горлом, адаптированной к диктору и акустическим условиям, и алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей применимы и к другим задачам распознавания речи, в том числе для других языков. В частности, в диссертации показана их эффективность в задаче распознавания английской спонтанной речи.
3. Результаты диссертации могут быть использованы при создании систем распознавания спонтанной речи для других языков, для которых отсутствуют большие обучающие базы. В этом случае следует обучать акустические модели для целевого языка, используя признаки, извлекаемые из глубокой нейронной сети с узким горлом, обученной по русским данным. Такой подход к построению акустических моделей позволяет значительно повысить точность распознавания, по сравнению с обучением акустических моделей только по малому количеству данных на целевом языке.

Перспективы дальнейшей разработки темы таковы:

1. Улучшение метода построения информативных признаков, извлекаемых из адаптированной к диктору и акустическим условиям глубокой нейронной сети, за счет обучения глубокой нейронной сети с узким горлом с использованием критериев разделения последовательностей.
2. Повышение точности распознавания русской телефонной спонтанной речи за счет применения акустических моделей на основе сверточных и рекуррентных нейронных сетей.
3. Повышение точности распознавания русской спонтанной речи с помощью применения подходов к построению языковых моделей, позволяющих эффективно учитывать дальний смысловой контекст, а также морфологическую, синтаксическую и семантическую информацию.
4. Повышение быстродействия системы распознавания русской телефонной спонтанной речи.

Список публикаций автора по теме диссертации

В журналах из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук

1. Меденников, И. П. Дикторо-зависимые признаки для распознавания спонтанной речи [Текст] / И. П. Меденников // Научно-технический вестник информационных технологий, механики и оптики. — 2016. — Т. 16., № 1. — С. 195–197.
2. Меденников, И. П. Двухэтапный алгоритм инициализации обучения акустических моделей на основе глубоких нейронных сетей [Текст] / И. П. Меденников // Научно-технический вестник информационных технологий, механики и оптики. — 2016. — Т. 16., № 2. — С. 379–381.

В изданиях, индексируемых в международных реферативных базах Scopus или Web of Science

3. Levin, K. Automated closed captioning for Russian live broadcasting [Text] / K. Levin, I. Ponomareva, A. Bulusheva, G. Chernykh, I. Medennikov, N. Merkin, A. Prudnikov, N. Tomashenko // Proc. Annual Conference of International Speech Communication Association (INTERSPEECH). — 2014. — P. 1438–1442.
4. Romanenko, A. Simplified Simultaneous Perturbation Stochastic Approximation for the Optimization of Free Decoding Parameters [Text] / A. Romanenko, A. Zatvornitsky, I. Medennikov // Speech and Computer, Lecture Notes in Computer Science. — 2014. — Vol. 8773. — P. 402–409.
5. Merkin, N. Controlling the Uncertainty Area in the Real Time LVCSR Application [Text] / N. Merkin, I. Medennikov, A. Romanenko, A. Zatvornitskiy // Speech and Computer, Lecture Notes in Computer Science. — 2014. — Vol. 8773. — P. 153–160.
6. Prudnikov, A. Improving Acoustic Models For Russian Spontaneous Speech Recognition [Text] / A. Prudnikov, I. Medennikov, V. Mendelev, M. Korenevsky, Y. Khokhlov // Speech and Computer, Lecture Notes in Computer Science. — 2015. — Vol. 9319. — P. 234–242.
7. Khomitsevich, O. A Bilingual Kazakh-Russian System for Automatic Speech Recognition and Synthesis [Text] / O. Khomitsevich, V. Mendelev, N. Tomashenko, S. Rybin, I. Medennikov, S. Kudubayeva // Speech and Computer, Lecture Notes in Computer Science. — 2015. — Vol. 9319. — P. 25–33.