Санкт-Петербургский институт информатики и автоматизации
Российской академии наук

На правах рукописи

КАРПОВ
Алексей Анатольевич

МОДЕЛИ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ НА ОСНОВЕ МОРФЕМНОГО АНАЛИЗА

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель:

к.т.н. Ронжин А.Л.

Санкт-Петербург - 2007

СОДЕРЖАНИЕ

Введение

Актуальность темы диссертации. Вопросами автоматического распознавания речи ученые стали заниматься с момента появления первых компьютеров, поскольку текстовый командный интерфейс взаимодействия с ЭВМ не обеспечивал приемлемой скорости и естественности работы. За многие годы исследований был разработан широкий спектр методов и компьютерных программ, направленных на решение проблем распознавания речи.

Сегодня получены многообещающие результаты и созданы действующие коммерческие системы, в основном, для английского языка, а также испанского, французского, японского, китайского и арабских языков. Это во многом связано с экономическими и политическими аспектами развития речевых технологий. Например, английский язык является наиболее распространенным и поэтому инвестиции в развитие технологий для автоматизированной обработки английской речи окупились достаточно быстро. В то же время речевым технологиям других языков уделяется недостаточно внимания, вследствие чего их развитие несколько сдерживается.

Между тем, русский язык является одним из самых популярных языков мира, на нем говорит свыше двадцати процентов населения Европы. Несмотря на это, действующих систем автоматического распознавания русской слитной речи фактически не существует. Кроме экономических проблем, на развитие Российских речевых технологий, в первую очередь, влияют особенности русского языка и речи, вызывающие сложности в процессе обработки. Основные из них: отсутствие строгих грамматических конструкций построения предложений, а также многочисленные правила словообразования, фонетического представления слов и расстановки ударений с большим количеством исключений.

Для оценки эффективности разрабатываемых систем автоматического распознавания речи применяют много показателей, интегральными же критериями оценки производительности таких систем служат точность распознавания речи (звуков, слов или фраз) и скорость обработки речевого сигнала. В идеальном случае система должна обеспечивать практически 100% точность распознавания речи при мгновенном выводе результата. Тем не менее,

учитывая ограниченные возможности существующих вычислительных ресурсов при решении таких сложных интеллектуальных задач как автоматическое распознавание речи человека, приходится находить компромисс между точностью и скоростью обработки.

Для улучшения характеристик распознавания русской слитной речи (в первую очередь скорости обработки), в диссертации предложен дополнительный морфемный уровень описания языка и речи, который вводится в каждый из двух этапов функционирования модели: обучение и распознавание. При этом за счет декомпозиции слов на морфемы обеспечивается акустико-лексическое моделирование большого количества словоформ языка при существенном сокращении размера словаря распознавания, что позволяет улучшить производительность и другие характеристики модели распознавания русской речи. Данный подход может быть также успешно использован и для других синтетических языков, например языков славянской группы (чешский, польский, украинский, белорусский, и т.д.), имеющих сходные с русским языком механизмы словообразования.

Цель работы и задачи исследования. Основной целью диссертационной работы является разработка модели дикторонезависимого распознавания русской слитной речи с большим словарем, которая обеспечивает ускорение процесса обработки речи при сохранении точности распознавания. Для достижения поставленной цели в ходе диссертационной работы поставлены и решены следующие задачи:

1. Анализ подходов к распознаванию английской и русской речи.
2. Выбор языковых единиц, наиболее подходящих для распознавания русской речи с большим словарем.
3. Разработка методов для модели обучения распознавателя русской речи, учитывающего специфику и морфологические особенности русского языка.

4. Разработка методов для модели распознавания русской речи с морфемным уровнем обработки языка и речи.

Методы исследования. Для решения поставленных задач в работе используются методы теории информации, теории множеств, теории вероятности, экспертного и статистического анализа. Компьютерная реализация разработанных алгоритмов производилась на основе объектно-ориентированного подхода.

Научная новизна работы состоит в следующем:

1. Разработана модель обучения распознавателя русской речи, включающая блоки создания акустических моделей русских фонем, декомпозиции словоформ языка и создания словаря лексических единиц с размером меньше слова (морфем), а также морфемной статистической модели языка прикладной области.
2. Разработана модель автоматического распознавания слитной русской речи, учитывающая механизмы словообразования и морфологические характеристики лексики русского языка и позволяющая существенно ускорить процесс распознавания за счет сокращения размера словаря.
3. Создана программная реализация распознавателя русской слитной речи с большим словарем SIRIUS, содержащая морфемный уровень обработки.

Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечивается за счет тщательного анализа состояния исследований в данной области, подтверждается корректностью предложенных моделей, алгоритмов и согласованностью результатов, полученных при компьютерной реализации, а также апробацией основных теоретических положений диссертации в печатных трудах и докладах на международных научных конференциях.

Практическая ценность работы. Разработанные модели, алгоритмы и программные реализации направлены на разрешение проблемы распознавания русской речи с большим словарем, возникающую из-за богатых механизмов словообразования и морфологических характеристик русского языка.

Предложенная модель обучения морфемного распознавателя русской речи позволяет на выходе получить три основных компонента, необходимых для функционирования модели распознавания: морфемный распознаваемый словарь с соответствующими фонетическими транскрипциями, статистическую морфемную модель языка предметной области и модели акустических фонетических единиц речи. Причем, за счет введения в модель уровня морфемного представления сокращается размер распознаваемого словаря, так как в процессе словообразования часто используются одни и те же морфемы, а также обеспечивается более полное покрытие пространства слов языка.

Разработанная модель распознавания речи, основанная на стохастическом моделировании речи посредством скрытых Марковских моделей и статистическом n-граммном моделировании языка предметной области с введением дополнительного уровня морфемного распознавания и синтеза слов, позволяет производить процесс распознавания с приемлемой точностью, обеспечив при этом значительный рост скорости обработки и устойчивости к синтаксическим отклонениям произнесенной фразы в ходе человеко-машинного взаимодействия по сравнению с традиционной целословной моделью распознавания речи.

Реализация результатов работы. Исследования, отраженные в диссертации, проведены в рамках научно-исследовательских работ: EC FP6-IST-2002-507609 SIMILAR Network of Excellence «The European taskforce creating human-machine interfaces SIMILAR to human-human communication»; INTAS № 04-77-7404 «Development of multi-voice and multi-language Text-to-Speech (TTS) and Speech-to-Text (STT) conversion system (languages: Belarussian, Polish, Russian)» и INTAS № 05-1000007-426 «Introduction of the automatic Russian speech recognition system SIRIUS in telecommunications»; ОИТВС РАН № 4.2 «Разработка методов статистической обработки речи для дикторонезависимых инфотелекоммуникационных приложений»; СПб НЦ РАН № М06-2.1К-29 «Разработка методов и программного обеспечения для дикторонезависимого распознавания русской речи с большим словарем». Кроме того, результаты работы использованы при разработке средств

голосового доступа к рубрикатору электронного каталога «Желтые Страницы» в рамках инновационного проекта совместно с компанией «NewVoice».

Апробация результатов работы. Основные положения и результаты диссертационной работы представлялись на Международных конференциях: «Interspeech – ICSLP 2006» (США, 2006); «Европейской конференции по обработке сигналов» EUSIPCO (Италия, 2006; Турция, 2005); «Intelligent Information Processing and Web Mining» (Польша, 2005); «Интеллектуальные многопроцессорные системы. Искусственный интеллект» (Украина, 2006; Россия, 2005); «Речь и Компьютер» SPECOM (Санкт-Петербург, 2006, 2004, 2002; Греция, 2005; Москва, 2003).

Публикации. Основные результаты по материалам диссертационной работы опубликованы в 24 печатных работах, в том числе в трех журналах ВАК («Известия ТРТУ», «Известия вузов. Приборостроение» и «Pattern Recognition and Image Analysis») и монографии серии «Информатика: неограниченные возможности и возможные ограничения» издательства «Наука».

Структура и объем работы. Диссертация объемом 129 машинописных страниц, содержит введение, четыре главы и заключение, список литературы (121 наименование), 8 таблиц, 44 рисунка.

Основные результаты. В ходе исследований, представленных в диссертации, были получены следующие основные результаты:

1. Разработана модель обучения распознавателя русской слитной речи с большим словарем с включением морфемного уровня анализа русского языка, что позволяет значительно сократить распознаваемый словарь языковых единиц и ускорить, таким образом, дальнейшую автоматическую обработку речи.
2. Разработана модель автоматического распознавания русской слитной речи с морфемным уровнем обработки языка и речи, использующая предложенный метод композиции слов из цепочек распознанных морфем.
3. Произведено сравнение реализованных целословной, морфемной, слоговой и фонемной моделей автоматического распознавания русской

речи по критериям точности распознавания на уровне фонем, слов и фраз, а также времени распознавания фраз. Выявлен существенный рост скорости обработки для морфемной модели при незначительном падении точности распознавания.

4. Создана программная реализация дикторонезависимого распознавания русской слитной речи, включающая в себя модели обучения и распознавания русской речи, а также модули для записи речевых данных и анализа гипотез и результатов распознавания речи.
5. Разработанные в диссертации методы и модели были реализованы в диалоговой модели голосового доступа к электронному справочному каталогу и многомодальной модели бесконтактного управления персональным компьютером.

Таким образом, в результате проведенных автором исследований предложено решение актуальной проблемы повышения производительности модели автоматического распознавания русской речи с большим словарем, что способствует внедрению технологий автоматического распознавания русской речи в различные области человеко-машинного взаимодействия.

В первой главе диссертации представлен анализ состояния дел в области автоматического распознавания речи в целом, а также сделан обзор существующих систем для русской речи, начиная с советских времен по настоящее время. Отмечается, что сейчас не существует готовых к использованию систем дикторонезависимого распознавания русской речи с большим словарем.

Во второй главе приводится описание модели обучения распознавателя русской речи с применением морфемного представления словаря и модели языка. Это позволяет значительно сократить размер словаря языковых единиц и улучшить производительность модели. Приведено сравнение моделей языка, основанных на различных языковых единицах (словах, морфемах, слогах) по размеру словаря распознавания и проценту непокрытых слов в тестовом

материале. Получены результаты, показывающие преимущество слоговой и морфемной моделей языка перед целословной моделью.

Третья глава описывает разработанную автором модель распознавания русской слитной речи с большим словарем с применением морфемного уровня обработки. Приведено описание и оценка предложенных методов параметрического спектрально-разностного представления речевого сигнала и выделения полезной речи методом анализа функции энтропии спектра звукового сигнала. Приводится результат сравнения целословной, морфемной, слоговой и фонемной моделей распознавания речи по двум основным критериям: точности и времени распознавания.

В четвертой главе приводятся данные по реализации разработанных алгоритмов в экспериментально-исследовательских моделях, использующих слитный ввод русской речи. Модели обучения и распознавания русской речи, а также модули для записи речевых баз данных и оценки результатов распознавания объединены в единую программную реализацию дикторонезависимого распознавания русской слитной речи SIRIUS. Эта программная реализация использована для создания модели голосового доступа к рубрикатору электронного справочного каталога «Желтые страницы», а также модели бесконтактного управления компьютером, где вместо клавиатуры и мыши для управления компьютером используется голосовой ввод и движения головы оператора.

Положения, выносимые на защиту:

1. Модель обучения распознавателя русской речи с морфемным представлением распознаваемого словаря и модели языка.
2. Модель распознавания русской речи с морфемным уровнем обработки языка и речи.
3. Программная реализация модели дикторонезависимого распознавания русской слитной речи с большим словарем.

Глава 1. Анализ требований и подходов к автоматическому распознаванию речи

Процесс автоматического распознавания речи представляет собой преобразование акустического сигнала, полученного от микрофона, в последовательность слов, которая затем может использоваться для понимания смысла речевого высказывания.

В первой главе диссертации приведен анализ основных требований, предъявляемых к системам распознавания речи, представлена базовая архитектура системы автоматического распознавания речи, опирающаяся на стохастические модели, а также сделан обзор существующих моделей распознавания русской речи.

1.1. Основные требования к современным системам распознавания речи

Задача распознавания речи характеризуется многими параметрами, в первую очередь, это свойства канала передачи речи, размер словаря, вариативность речи, уровень окружающего шума, тип ввода речи (изолированная/слитная) [37].

Для распознавания изолированных слов необходимо, чтобы диктор делал короткие паузы между словами, что замедляет ввод и ухудшает естественность, в то время как при вводе слитной речи этого не требуется. В отличие от печатного текста или от искусственных сигналов естественная речь не допускает простого и однозначного членения на элементы (фонемы, слова, фразы), поскольку эти элементы не имеют явных физических границ. Они вычленяются в сознании слушателя – носителя данного языка в результате сложного многоуровневого процесса распознавания и понимания речи [66]. Если попросить слушателя записать в виде фонем незнакомую иностранную речь, то он совершит множество ошибок членения слов и фраз, т.е. даже человек не может членить речь без использования знаний лексики, грамматики,

смысла. Границы слов могут быть определены лишь в процессе распознавания, посредством подбора оптимальной последовательности слов, наилучшим образом согласующейся с входным потоком речи по акустическим и лингвистическим критериям.

Сложность проблемы распознавания речи, главным образом, связана с вариативностью ее основных параметров, на которые влияет множество факторов. Прежде всего, это случайная компонента процесса речеобразования, которая приводит к многообразию описаний одного и того же слова, произнесенного одним и тем же диктором. Более существенная вариативность связана с индивидуальными различиями речевых аппаратов разных дикторов. Тут нужно также отметить влияние пола диктора, возрастных различий, диалектов, эмоционального и физического состояния диктора. Кроме того, значительное влияние вносит акустический аспект, т.е. смена микрофона, расположение его относительно рта, акустическая обстановка в помещении.

Точность распознавания существенно ухудшается с увеличением словаря, так как при этом, появляются группы акустически подобных слов, что приводит к акустической неоднозначности, причем она экспоненциально усиливается с ростом словаря. Существует несколько возможных классификаций размера распознаваемого словаря. Согласно [9] малым словарем считается словарь, содержащий единицы и десятки слов. Задач и приложений, где используется малый словарь распознавания, очень много: распознавание последовательностей цифр (номеров телефонов) [78]; системы речевого командного управления подвижными техническими объектами (автомобилем, самолетом, и т.д.) [94], системы дистанционного управления роботами [63, 31], системы управления оборудованием (например, медицинским) [112] и т.д. Средний распознаваемый словарь содержит сотни слов. Такого словаря достаточно для большинства диалоговых или запросно-ответных систем [22, 114]. Большой словарь начиняется от тысяч слов [9], такие системы распознавания могут использоваться в автоматизированных справочных системах или системах диктовки в ограниченной предметной области. Словарь

размером свыше сотни тысяч слов считается сверхбольшим [86] и он позволяет реализовывать системы стенографии практически любого текста (для аналитических языков).

При работе с реальной диалоговой системой или при вводе текста голосом пользователь хочет получить ответ от системы незамедлительно, он не готов ждать даже несколько секунд, поэтому система, распознающая речь должна работать в режиме реального времени без существенных задержек в ответе. Конечно, существуют задачи распознавания, где время реакции не играет существенной роли, например преобразование в текст архивных звуковых записей [105], но число таких приложений очень невелико.

Таким образом, наиболее важными требованиями, которым должны стремиться удовлетворить современные системы автоматического распознавания речи, являются: слитный ввод речи, дикторонезависимость, способность распознавать большое количество слов и высокое быстродействие системы.

Крайне важной задачей является многокритериальное оценивание таких сложных интеллектуальных систем, как системы распознавания речи, и обоснованный выбор оптимальных моделей и их параметров [104, 62]. Для оценки эффективности разрабатываемых систем автоматического распознавания речи применяют целый ряд критериев на каждом из уровней обработки речи, среди них два критерия являются интегральными: точность распознавания и время реакции (ответа) системы. Идеальная автоматическая система должна мгновенно выдавать безошибочный результат. Компьютерные системы в ближайшие годы не смогут достичь таких показателей, но должны стремиться к производительности биологического прототипа (человека).

1.2. Базовые подходы к автоматическому распознаванию речи

Задача распознавания речи состоит в подборе оптимальной последовательности моделей слов, которая наиболее вероятна (правдоподобна)

обрабатываемому речевого сигналу. Анализ обзорных статей ведущих мировых ученых [84, 83, 79, 113] показал, что в настоящее время практически все системы автоматического распознавания речи строятся на основе нескольких базовых подходов (рисунок 1.1): скрытые Марковские модели, искусственные нейронные сети, динамическое программирование.

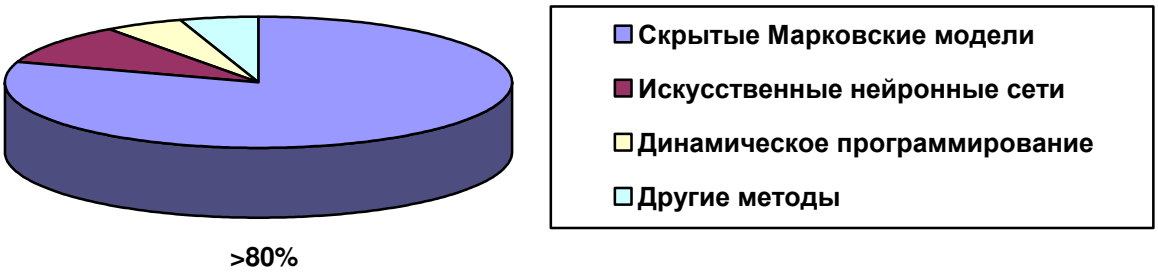

Рис. 1.1. Базовые подходы к автоматическому распознаванию речи

Долгое время подход на основе динамического программирования (ДП) был доминирующим. Он позволяет производить сравнение речевого фрагмента с созданным заранее эталоном слова. Для того чтобы сравнить слово с эталоном, надо путем деформации оси времени совместить участки, соответствующие одним и тем же звукам, измерить остаточные различия между ними и просуммировать эти частные расстояния, взятые с некоторыми весовыми коэффициентами. Задача ДП сводится к поиску оптимального нелинейного согласования двух отрезков речи. Для этого широко использовались алгоритмы ДП, базирующиеся на фундаментальных работах Р. Беллмана [4]. Одна из первых публикаций по применению ДП в распознавании речи принадлежит украинскому ученому Т.К. Винцюку [8]. Существует несколько подходов к распознаванию слитной речи методами ДП: двухуровневый алгоритм динамического программирования, метод построения уровней (level-building) и однопроходный (one-pass) метод [106]. Алгоритмы используют одинаковые базовые принципы и отличаются вычислительной сложностью, объемом памяти и сложностью реализации. Недавно был также

предложен метод распознавания слитной речи на основе ДП с применением анализа речи в скользящем окне и теории размытых множеств [36].

Основным недостатком подходов, основанных на ДП, является их дикторозависимость. Кроме того, каждый новый пользователь системы, перед тем как ее использовать, должен создать свои эталоны, т.е. наговорить все слова, которые присутствуют в словаре. Для повышения надежности распознавания при записи эталонов пользователю приходится повторять все слова по несколько раз. По этой причине такой подход сейчас используется лишь для приложений с малым словарем, например, вызов определенного абонента в мобильных телефонах или персонифицированное голосовое управление офисными программами.

Искусственные нейронные сети (ИНС) также используются при распознавании речи. Они представляют собой попытку использования процессов, происходящих в нервных системах биологических организмов. При правильно выбранной структуре сеть, натренированная на определенном наборе обучающих выборок, будет выдавать правильные результаты при подаче на ее вход данных, относящихся к тому же множеству, но непосредственно не участвующих в процессе обучения. На практике используются нейронные сети, имеющие один или несколько скрытых слоев нейронов между входом и выходом сети [29, 10]. В этом случае сложность сети определяется количеством нейронов в скрытом слое, так как количество нейронов во входном и выходном слоях фиксировано и зависит от условий задачи. Распространенным является подход, когда на входы нейронной сети подаются вектора признаков речевого сигнала, а выходы сети связаны с распознаваемым словарем (количество выходов равняется количеству слов в словаре). Нейронные сети способны обучаться на голосах нескольких дикторов, позволяя создавать дикторонезависимые системы распознавания, однако их применение для слитной речи затруднительно, так как при слитном вводе неизвестна заранее длительность речевого сигнала, а соответственно и количество векторов признаков, а также количество и порядок произнесенных

слов, что значительно затрудняет создание и обучение сети. Однако нейронные сети иногда применяют в комбинированных со скрытыми Марковскими моделями системах распознавания речи [30]. В этом случае нейронные сети задействуются либо на уровне предобработки векторов признаков речи, либо на уровне постобработки текстов гипотез распознавания. Несмотря на высокий потенциал, ИНС в области распознавания речи пока не получили широкого применения, поскольку их обучение имеет большую сложность и требует больших вычислительных ресурсов.

В настоящее время наиболее популярным математическим аппаратом для автоматического распознавания речи являются скрытые Марковские модели (СММ) [25, 15]. Они довольно содержательны по своей математической структуре, поэтому стали теоретическим фундаментом для различных областей исследований случайных процессов, не только речи [26]. СММ позволяют решать задачи распознавания речи, а также улучшать качество сигнала, загрязненного шумами и искажениями, моделировать источник речевого сигнала, оптимизировать структуру диалога и др. Сейчас подавляющее большинство систем распознавания речи строится на основе СММ, так как для них предложены достаточно эффективные методы дикторонезависимого распознавания слитной речи.

К остальным технологиям, которые также исследуются для решения задачи автоматического распознавания речи можно отнести: Support Vector Machines [110], вейвлет анализ речи [11] и системы моделирования человеческого уха. Однако данные технологии не находят массового применения в современных системах распознавания речи.

1.3. Распознавание речи на основе вероятностных моделей

На рисунке 1.2 показана общая схема распознавателя речи, построенного с использованием аппарата СММ [106]. Человек произносит некоторую фразу, которая представляет собой последовательность слов $W = w_1, \ldots, w_N$. Задача системы распознавания речи заключается в том, чтобы правильно распознать

эту последовательность слов. Однако в ходе распознавания могут возникать ошибки, поэтому результат может оказаться отличным от W, например $W' = w'_1, \ldots, w'_M$. Для параметрического описания речевого сигнала, он разделяется на короткие сегменты, которые затем преобразуются в вектора признаков $O = o_1, \ldots, o_T$.

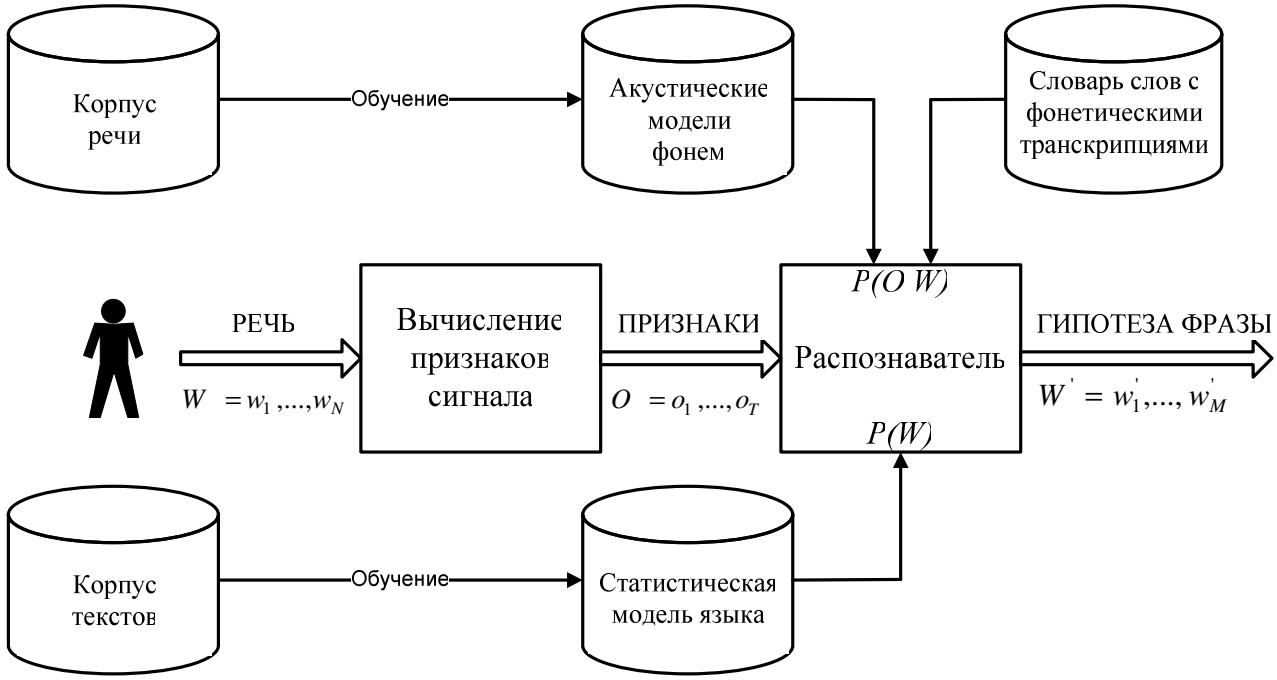

Рис. 1.2. Базовая архитектура системы дикторонезависимого распознавания слитной речи

Вычисленные вектора признаков и поступают в распознаватель речи, задача которого заключается в выборе гипотезы фразы с максимальной вероятностью, которая задается формулой Байеса [106]:

$$W' = \arg\max_W P(W|O) = \arg\max_W \frac{P(W)\cdot P(O|W)}{P(O)} = \arg\max_W P(W)\cdot P(O|W) \qquad (1.1)$$

где $P(W)$ и $P(O|W)$ – вероятностные оценки модели языка и акустической модели речи соответственно. Эти модели создаются заранее в ходе процесса обучения, где входными данными являются текстовые и речевые корпуса. Формула 1.1 определяет вероятность наступления события в условиях, когда на

основе наблюдений известна лишь некоторая частичная информация о событиях.

Далее последовательно по уровням обработки сигнала описывается общепринятая архитектура дикторонезависимого распознавания слитной речи, использующая вероятностные и статистические методы моделирования речевого сигнала и языка предметной области.

1.3.1. Признаковое описание речевого сигнала

Важным вопросом, с которым в первую очередь сталкивается разработчик речевых технологий – это разработка оптимального метода параметрического представления речевого сигнала, который позволил бы достаточно хорошо различать звуки и слова речи и в то же время обеспечивать инвариантность к особенностям произношения дикторов и к изменениям акустической обстановки. Из практики известно, что большинство ошибок распознавания слов происходит по причине вариаций громкости сигнала, которые происходят или из-за неточной фиксации микрофона, либо изменения громкости произношения. Другой частой причиной ошибок являются случайные нелинейные деформации формы спектра, которые постоянно присутствуют в речевом сигнале у одного и того же диктора. Поэтому одной из важнейших задач при создании робастных систем распознавания речи является выбор такого представления анализируемого сигнала, которое является достаточно адекватным его содержанию и в то же время инвариантным к голосам дикторов и различным акустическим обстановкам.

На практике, речевой сигнал, полученный с микрофона, оцифровывается с частотой дискретизации 8–22 кГц. Последовательность цифровых отсчетов разделяется на сегменты речи длительностью 10–30 мс, такая длительность примерно соответствует квазистационарным участкам речи. Для каждого сегмента вычисляется вектор параметров (признаков); эти вектора затем используется на акустическом уровне распознавания речи.

Сейчас существует широкий спектр методов параметрического представления сигнала на основе автокорреляционного анализа, аппаратной полосовой фильтрации, методов спектрального анализа, коэффициентов линейного предсказания (КЛП) [24, 21]. Однако наиболее распространены два класса методов, использующие разновидности спектрального анализа и КЛП. Популярность этих методов объясняется тем, что они хорошо согласуются с моделями слухового восприятия и речеобразования человека соответственно.

Техника линейного предсказания, основанная на авторегрессионном анализе, широко распространена в системах сжатия речи. Основной принцип линейного предсказания основан на аппроксимации текущего отсчета речевого сигнала через линейную комбинацию соседних отсчетов. При линейном предсказании основная задача анализа речи - найти коэффициенты этой линейной комбинации, которые дают минимальную ошибку предсказания на участке анализа сигнала. Коэффициенты предсказания - это весовые коэффициенты, используемые в линейной комбинации, которые вычисляются путем минимизации среднего квадрата разности между отсчетами речевого сигнала и их предсказанными значениями.

Самым распространенным подходом к параметризации речи является спектральный анализ сегментов сигнала с вычислением их кепстральных признаков [57, 109]. Изучение спектрального преобразования сигнала привели многих исследователей к весьма сходным алгоритмам вычисления оптимального набора параметров. Эти алгоритмы включают в себя: предварительную обработку цифровых отсчетов, например, предыскажающий фильтр или процедуру весовой обработки окна; быстрое преобразование Фурье (БПФ) над сегментами речи длиной 10–30 мс.; формирование набора перекрывающихся фильтров, расположенных эквидистантно или по некоторому нелинейному закону, например, согласно *Mel* или *Bark* шкале [35]. Отсчеты БПФ, включенные в каждый фильтр, пересчитываются с учетом треугольного окна, затем определяется интегральная энергия $S(n)$ на выходе каждого фильтра и далее производится логарифмирование выхода каждого

фильтра. Этот набор данных подвергается косинус-преобразованию, что в итоге приводит к кепстральным коэффициентам C_{ij}. Кепстральные коэффициенты, полученные с использованием Mel-шкалы треугольных фильтров, называются Mel-частотными кепстральными коэффициентами.

$$C_{ij} = \sum_{n=1}^{N} \log(S(n)) \cos\left[j(n - \frac{1}{2}) \frac{\pi}{N} \right], \quad 1 \le j \le J , \tag{1.2}$$

где j - номер кепстрального коэффициента.

Таким образом, в результате процедуры параметрического представления речевой сигнал преобразуется в последовательность векторов признаков, после переходит на более высокий уровень моделирования – звуков и слов.

1.3.2. Акустико-фонетическое моделирование речи

В настоящее время наиболее эффективным аппаратом для моделирования и распознавания естественной речи считаются скрытые Марковские модели [15, 100]. Теоретическая база математических моделей разработана петербургским профессором А.А. Марковым в начале XX века [25]. Сейчас методами Марковского моделирования пользуются многие исследователи случайных процессов. Разновидность Марковских моделей, называемая скрытая Марковская модель (СММ), основанная на теории дискретных случайных цепей, была впервые введена и изучена в конце 60-х – начале 70-х годов. СММ – это дважды стохастический процесс. Термин «дважды» используется для обозначения такой пары процессов, один из которых является основным, но скрытым от нас и наблюдаемым только через другой стохастический процесс.

Модели такого типа особенно удобны для описания речевого сигнала, поскольку в действительности давление звуковой волны, которое мы измеряем, представляет собой только некоторый код основного символьного процесса, протекающего в ненаблюдаемых и полностью недоступных участках мозга [15]. В наблюдаемом акустическом процессе выявляются измеримые физические корреляты лингвистической структуры.

При построении модели распознавания речи на основе СММ выбирают их основные параметры: тип модели (эргодическая, модель Бэкиса, лево-правая модель и др.), размер модели (число состояний), тип наблюдаемых параметров (дискретные или непрерывные плотности распределения векторов наблюдений).

Если распознаваемый словарь небольшой, то можно для каждого слова создать вручную топологию СММ. При этом обычно количество состояний модели равняется количеству звуков (фонем) в слове. Для распознавания речи с большим словарем крайне затруднительно построить и обучить индивидуальную СММ для каждого слова. Поэтому каждое слово преобразуется в последовательность произносимых фонем, и строится модель для каждой фонемы. Как правило, модель фонемы имеет 3 состояния: первое описывает начало фонемы, второе представляет центральную часть фонемы и третье – окончание фонемы (рисунок 1.3). На рисунках кружками обозначены состояния моделей, стрелками – переходы между ними. Также модели могут строиться не только для фонем, но и для аллофонов (фонем в акустическом контексте соседних фонем) и иных акустических единиц речи, таких как полуслоги, дифоны, полуаллофоны и т.д. [67].

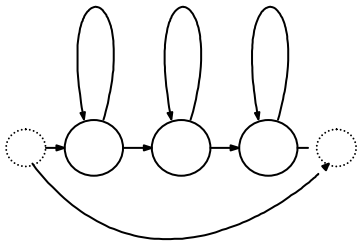

Рис. 1.3. Модель фонетической единицы речи

СММ слова получается путем соединения в цепочку моделей фонем из соответствующего фонетического алфавита, что представлено на рисунке 1.4. Аналогичным образом соединяются модели слов друг с другом, образуя модели фраз и предложений.

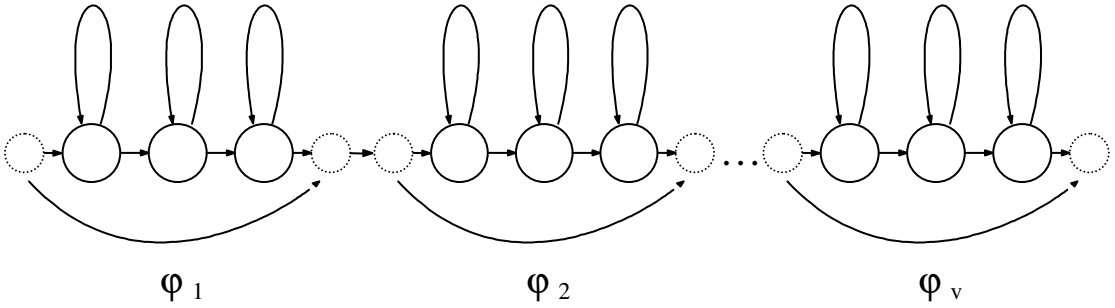

Рис. 1.4. СММ слова, содержащая фонетические элементы из алфавита φ

На рисунке 1.5 приведен пример моделирования слова «солнце». В данном случае модель позволяет представить два варианта произнесения данного слова, когда каждый звук проговаривается и когда фонема [л] пропадает, причем вероятность пропадания данной фонемы оценивается в 80%, а вероятность произнесения – в 20%. Для всех остальных фонем слова вероятность пропадания равна 0.

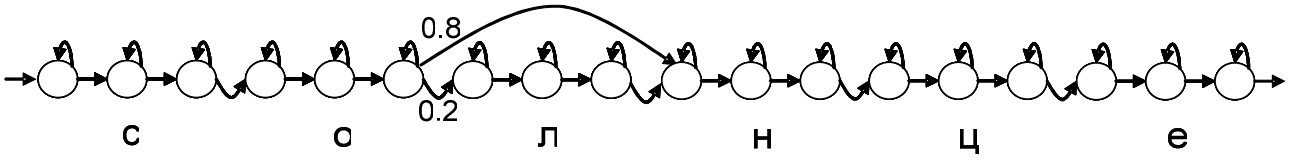

Рис. 1.5. Вид СММ, описывающей слово «солнце»

Каждому состоянию соответствует набор векторов параметров, наблюдаемых в этом состоянии, с вероятностью их наблюдения. Для полного определения скрытой Марковской модели $\lambda = (N, M, A, B, \pi)$ некоторой речевой единицы необходимо задать следующие параметры [106]:

1) *N*, число состояний в модели. Хотя состояния скрыты от наблюдателя, в практических задачах состояниям или множествам состояний модели приписывается определенный физический смысл (например, фонема). Для обозначения множества состояний всей модели

используется запись $S = \{S_1, S_2, ...S_N\}$, а текущее состояние модели в момент времени t обозначается q_t.

2) M, число различных символов наблюдения, которые могут порождаться моделью, т. е. размер дискретного алфавита. Символы наблюдения соответствуют физическому выходу моделируемой системы. Все множество допустимых наблюдаемых символов (векторов) обозначается как $V = \{v_1, v_2, ...v_M\}$. Этот параметр необходим только для дискретных СММ, для непрерывных моделей он не применяется.
3) Распределение вероятностей переходов между состояниями (матрица переходных вероятностей) $A = \{a_{ij}\}$,

   где $a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i]$.
4) Распределение вероятностей появления символов наблюдения в состоянии j, $B = \{b_j(k)\}$,

   где $b_j(k) = P[o_t = v_k \mid q_t = S_j]$, причем o и v являются векторами наблюдений.
5) Вероятностное распределение начальных состояний $\pi = \{\pi_i\}$ модели,

   где $\pi_i = P[q_1 = S_i]$.

Проблема, возникающая при использовании дискретных Марковских моделей, заключается в том, что в большинстве практических задач наблюдения являются непрерывными сигналами (или векторами) и их квантование с помощью кодовых книг (размерностью от 16 до 512) может иногда приводить к серьезным искажениям исходного сигнала. Поэтому часто для распознавания речи используют СММ с непрерывными плотностями наблюдений. В таких моделях плотность вероятности векторов наблюдений описывается следующим образом:

$$b_j(O) = \sum_{m=1}^{M} C_{jm}\, \vartheta[O, \mu_{jm}, U_{jm}], \qquad (1.3)$$

где O - моделируемый вектор наблюдений, C_{jm} - весовой коэффициент m-ой компоненты в состоянии j и ϑ - произвольная логарифмически-вогнутая или эллиптически-симметричная плотность вероятности (например, Гауссова плотность вероятности) с вектором средних значений μ_{jm} и ковариационной матрицей U_{jm} для m-й составляющей смеси в состоянии j. Как правило, в качестве плотности вероятности используется Гауссова плотность. Плотности такого вида часто используются на практике, поскольку позволяют с любой точностью аппроксимировать произвольную непрерывную функцию плотности вероятности, содержащую конечное число компонент.

Существует три основные задачи, связанные с использованием СММ [106]:

Задача 1 (оценка вероятности). Пусть задана последовательность наблюдений $O = O_1 O_2 ... O_T$ и модель λ. Как эффективно вычислить величину $P(O \mid \lambda)$, т. е. вероятность появлений этой последовательности наблюдений для данной модели? Существует несколько способов оценки правдоподобия. Наиболее широко применяют алгоритмы прямого-обратного хода (forward-backward), а также лучевые алгоритмы [33].

Задача 2 (декодирование). Пусть заданы последовательность наблюдений $O = O_1 O_2 ... O_T$ и модель λ. Как выбрать последовательность состояний $Q = q_1 q_2 ... q_T$, которая в некотором значимом смысле будет оптимальной (например, наилучшим образом соответствует имеющейся последовательности наблюдений)? Для решения этой задачи применяют алгоритм Витерби [116, 33].

Задача 3 (обучение). Каким образом нужно подстроить параметры модели λ, для того чтобы максимизировать $P(O \mid \lambda)$? Задача обучения СММ является крайне важной и в то же время наиболее трудной задачей.

Цель обучения акустических моделей состоит в том, чтобы по заданной последовательности наблюдений определить метод такой подстройки параметров модели, чтобы для полученной модифицированной модели вероятность появления этой последовательности была максимальной. Не

существует известного аналитического выражения для настройки параметров такой модели. Кроме того, на практике, располагая некоторой последовательностью наблюдений в качестве обучающих данных, нельзя указать оптимальный способ оценки параметров. Тем не менее, используя итеративные процедуры, например метод Баума-Уэлча, EM-метод или градиентные методы [76, 107, 33], можно выбрать параметры модели таким образом, чтобы локально максимизировать вероятность $P(O \mid \lambda)$.

Если итеративно повторять процедуру переоценки параметров, используя на каждом новом шаге значения параметров модели, полученные на предыдущем шаге, то будем последовательно получать модели, для которых вероятность появления последовательности наблюдений O будет увеличиваться. Процедура продолжается до тех пор, пока не будет достигнута некоторая предельная точка (например, по критерию максимума правдоподобия СММ).

Согласно теории [106], процедура переоценки должна давать значения параметров СММ, которые соответствуют локальному максимуму функции правдоподобия. И при этом крайне важным является вопрос, как выбирать начальные значения параметров заданной модели, для того чтобы локальный максимум оказался глобальным максимумом функции правдоподобия.

Исследования показывают [106], что либо случайные (подверженные стохастичности и ограничениям ненулевых значений), либо однородные начальные оценки параметров π и A почти во всех случаях позволяют получать вполне приемлемые повторные оценки для этих параметров. Что же касается параметра B, то хорошие начальные оценки являются полезными в случае дискретных символов и необходимы в случае непрерывного распределения. Такие начальные оценки могут получаться несколькими различными способами, включая ручную сегментацию последовательностей наблюдений на состояния с усреднением числа наблюдений в состояниях, сегментацию наблюдений по методу максимального правдоподобия с усреднением, сегментацию с использованием метода k-средних [90] и т.д.

После инициализации модели множество обучающих последовательностей наблюдений разбивается на состояния в соответствии с используемой моделью λ. Такое разбиение достигается посредством нахождения оптимальной последовательности состояний с помощью алгоритма Витерби и последующего поиска в обратном направлении вдоль оптимального пути. Результатом разбиения на состояния каждой обучающей последовательности также является вероятностная оценка принадлежности множества наблюдений конкретной модели. Обновленная модель $\bar{\lambda}$ получается на основе вычисленных параметров модели, а переоценка всех параметров этой модели выполняется с помощью процедуры повторного оценивания. Результирующая модель сравнивается с предыдущей моделью посредством вычисления меры отклонения, которая отражает статистическое сходство этих моделей. Если эта мера отклонения моделей превышает порог, старая модель λ заменяется новой моделью $\bar{\lambda}$ (для которой выполняется процедура переоценки), и полностью повторяется цикл обучения. Если же мера отклонения не превышает данного порога, то полагается, что модель сходится, и сохраняются параметры последней модели.

На этапе автоматического распознавания речи строятся всевозможные переходы по состояниям СММ и определяется вероятность того, что в конце мы окажемся в конечном состоянии, используя алгоритм прямого-обратного хода или алгоритм Витерби. Алгоритм Витерби применяют как для распознавания изолированной, так и слитной речи. Он состоит из прямого и обратного проходов и реализуется следующим образом [34]. Для начала необходимо ввести следующую переменную:

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P[q_1 q_2 \ldots q_{t-1}, q_t = S_i, o_1 o_2 \ldots o_t \mid \lambda] \qquad (1.4)$$

имеющую смысл максимальной вероятности того, что при заданных наблюдениях до момента времени t последовательность состояний завершится в момент времени t в состоянии S_i. Также введем переменную $\psi_t(j)$ для хранения аргументов, максимизирующих $\delta_t(j)$. Алгоритм состоит из 4 шагов:

1) Инициализация

$\delta_1(j) = \pi_i b_i(o_1)$, $1 \le i \le N$

$\psi_1(i) = 0$

2) Индуктивный переход

$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t)$, $1 \le i \le N$, $2 \le t \le T$

$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]$

3) Останов

$P^* = \max_{1 \le i \le N} [\delta_T(i)]$ Определяет максимальную вероятность наблюдения последовательности $o_1 o_2 \ldots o_T$, которая достигается при прохождении некой оптимальной последовательности состояний $Q^* = q_1^*, \ldots q_T^*$, для которой к настоящему моменту известно только последнее состояние:

$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$

4) Восстановление оптимальной последовательности состояний (обратный проход): $q_t^* = \psi_{t+1}(q_{t+1}^*)$, t = T-1, T-2, …, 1

Результатом работы алгоритма является наибольшая вероятность появления распознаваемой последовательности наблюдений для заданной СММ, то есть степень близости слова (или цепочки слов), задаваемого данной моделью, к распознаваемому сигналу. Причем с помощью алгоритма Витерби можно как вычислить вероятность принадлежности последовательности наблюдений некоторой СММ, так и узнать оптимальную последовательность пройденных состояний модели.

1.3.3. Статистическое моделирование языка предметной области

Неотъемлемым компонентом распознавателя слитной речи является модель языка, используемая при генерации гипотез фраз. Одной из самых популярных синтаксических моделей естественного языка являются N-граммы [97]. Эта модель была предложена в середине 1980-х гг. Ф. Джелинеком [14]

является статистической и ее цель состоит в оценке вероятности появления цепочки слов $W = (w_1 w_2 \ldots w_q)$ в некотором тексте.

N-грамма – это последовательность из n элементов (например, слов), а n-граммная модель языка используется для предсказания элемента в последовательности, содержащей $n-1$ предшественников. Эта модель основана на предположении, что вероятность какой-то определенной n-граммы, содержащейся в неизвестном тексте, можно оценить, зная, как часто она встречается в некотором обучающем тексте.

Вероятность $P(w_1, w_2, ..., w_m)$ можно представить в виде произведения условных вероятностей:

$$P(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} P(w_i \mid w_{1,} w_{2,} ... w_{i-1}) \qquad (1.5)$$

Отсюда видно, как можно легко аппроксимировать $P(W)$ при ограниченном контексте длиной $n-1$:

$$P(w_1, w_2, ..., w_m) \cong \prod_{i=1}^{m} P(w_i \mid w_{i-n+1,} w_{i-n+2,} ... w_{i-1}) \text{ , } n > 1 \qquad (1.6)$$

Оценить вероятность появления слова в зависимости от всей предыдущей цепочки пока что не представляется возможным вследствие вычислительной сложности задачи. Выбор значения n существенно влияет на количество возможных параметров, которыми должна обладать модель, чтобы максимально приблизиться к $\left|W^n\right|$, где $W*$ – это ряд слов в языковой модели из словаря. Так 3-граммная модель со словарем 100000 слов теоретически может содержать до $100000^4 = 10^{20}$ параметров. Поэтому в n-граммной модели языка вероятность появления каждого слова считается функцией от предшествующих $n-1$ слов и на практике используют значения n в пределах от 1 до 3.

Вероятностные значения n-граммных моделей основываются на максимальной вероятности событий, вычисленных в контексте обучающего текста. Так условная вероятность появления триграммы в тексте может быть вычислена следующим образом:

$$P(w_i \mid w_{i-1,} w_{i-2}) = \frac{C(w_{i-2,} w_{i-1,} w_i)}{C(w_{i-2,} w_{i-1})} \quad (1.7)$$

где C – количество наблюдений данной последовательности слов в тексте.

На практике обучающие данные всегда неполны, то есть значительная часть теоретически возможных n-грамм либо вообще отсутствуют, либо встречается слишком редко для того, чтобы можно было применить стохастические методы для оценки вероятности их появления. Если такая n-грамма встретится во время работы, то правильный вариант распознавания будет отклонен или его вероятность будет существенно занижена. Таким образом, не целесообразно принимать значения вероятностей всех необнаруженных последовательностей равным нулю. Поэтому были разработаны методы сглаживания вероятностных параметров языковых моделей [93].

Другой подход к построению языковых моделей основан на использовании формальных грамматических правил, описывающих корректные предложения языка [101]. Обычно правила для таких языковых моделей строятся «вручную» экспертом, что сопряжено со значительными трудностями. К сожалению, эти языковые модели плохо походят для обработки естественного языка, поскольку фразы, содержащие некоторые отклонения от правил, будут отвергнуты системой. Даже при распознавании английского языка с его строгой грамматикой и порядком слов в предложении, разработчики постепенно отказываются от этого подхода в сторону статистических n-граммных моделей. Основным достоинством статистических моделей языка является возможность автоматического построения модели по обучающему корпусу достаточно большого размера и относительно высокая скорость работы.

1.3.4. Метод распознавания слитной речи

Для работы со слитной речью необходимо соединить скрытые Марковские модели слов в одну общую СММ языка предметной области с

учетом вероятностей переходов между словами, которые задаются моделью языка. Каждая модель в последовательности напрямую связана с элементом, лежащим в ее основе. Этими элементами могут быть целые слова или части слов, такие как фонемы. На рисунке 1.6 показана сеть, в которой каждое слово определено как последовательность скрытых Марковских моделей, основанных на фонемах, и все слова замкнуты в петлю (цикл). В этой сети кружками показаны СММ, а прямоугольниками – состояния конца слова.

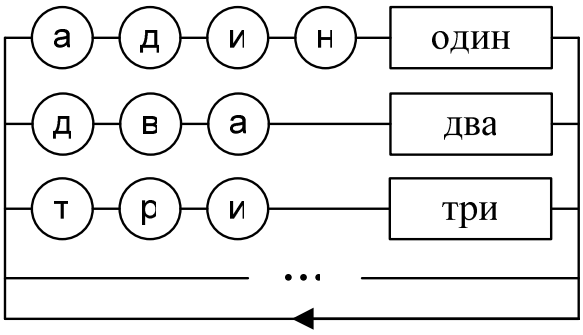

Рис. 1.6. Пример объединенной СММ для метода прохождения маркеров

Таким образом, распознающая сеть, в итоге, состоит из состояний скрытой Марковской модели, соединенных переходами. В ней можно выделить три различных уровня: слов, фонем и состояний. На рисунке 1.7 показана эта иерархия.

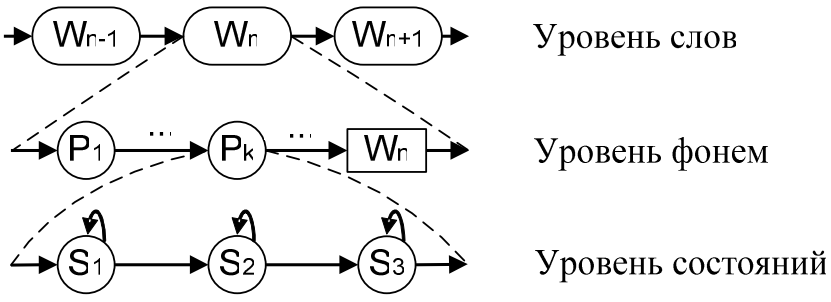

Рис. 1.7. Три уровня описания распознающей сети

Процедура Баума-Уэлча для обучения не требует особенной модификации при работе со слитной речью, однако алгоритм Витерби для распознавания требует усовершенствований, так как заранее неизвестны границы слов во фразе и их количество. Для распознавания слитной речи предложен модифицированный алгоритм Витерби, называемый метод передачи маркеров (token passing method) [119] и реализованный в инструментарии Hidden Markov Model Toolkit [43]. Метод передачи маркеров определяет прохождение возможных путей по состояниям объединенной СММ. В начало каждого слова ставится маркер и применяется итеративный алгоритм оптимизации Витерби, сдвигая маркер на каждом шаге и вычисляя для него акустическую вероятность. Предположим, в каждом состоянии *j* скрытой Марковской модели в момент времени *t* находится отдельный маркер, который содержит значение логарифма вероятности $\Psi_j(t)$ пройденной части пути. Этот маркер отображает соотношение между наблюдаемой последовательностью от o_1 до o_t и моделью, позволяющее заключить, что модель находится в состоянии *j* в момент времени *t*. Для вычисления вероятности на каждом шаге алгоритма используется рекурсивная формула:

$$\psi_j(t) = \max_i\{\psi_i(t-1) + \log(a_{ij})\} + \log(b_j(o_t)) \qquad (1.8)$$

Эта формула используется в алгоритме, который выполняется в каждый момент времени *t* для каждого маркера. Ключевые шаги алгоритма следующие:

1) Копия каждого маркера, находящегося в состоянии *i*, должна пройти через следующее состояние *j*, тогда приращение логарифма вероятности в маркере будет равняться $\log[\ a_{ij}\ ] + \log[\ b_j(o(t))]$.
2) Проверка маркеров в каждом состоянии и удаление всех маркеров, кроме маркеров с самой высокой вероятностью.

При достижении состояния конца некоторого слова в маркер записывается его индекс, а при выходе из каждого состояния маркеры размножаются (копированием) по числу дальнейших переходов в модели. При этом в маркер записывается его путь (история) через сеть. Когда маркер

переходит от выходного состояния одного слова к входному состоянию другого, переход представляет собой потенциальную границу слов, которая и записывается в историю маркера.

В итоге после обработки всей последовательности векторов наблюдений выбирается маркер, имеющий наибольшую вероятность. Когда наилучший маркер достигает конца обрабатываемого сигнала (последовательности наблюдений), то путь, которым он проходит через сеть, известен в виде истории (хранящейся в маркере) и из маркера считывается последовательность пройденных слов, которая и является гипотезой распознавания фразы. Данная методика распознавания слитной речи эффективно используется в настоящее время для автоматического распознавания речи на многих языках.

1.4. Обзор моделей распознавания русской речи

Распознаванием и анализом русской речи ученые и инженеры начали заниматься еще в 40-х годах прошлого века [27]. Первые исследования, в основном, имели военную направленность, затем все большую активность в данной области проявляли гражданские специалисты. Из наиболее известных систем распознавания русской речи можно привести устройства линии «Речь» [7], разработанные под руководством Т. Винцюка (Киев, Украина). В основе данной системы заложена концепция последовательной обработки речевой информации на основе динамического программирования и временном представлении речи как результата нелинейного сжатия и растяжения.

Иное направление в области распознавания речи было заложено В. Трунинским-Донским [65]. В его исследованиях особое внимание уделялось акустическим признакам способа и места образования речи (временные, частотные, амплитудные) для принятия решения на каждом шаге обработки речевой информации. В этом принципиальное отличие данного подхода от концепции, принятой в работах коллектива Винцюка, опирающейся на математический метод.

Некоторые успешные разработки систем распознавания были доведены до опытно-конструкторских работ и даже запущены в серийное производство. Устройства распознавания-синтеза речи МАРС-1, МАРС-2 [2, в] основаны на формантном методе анализа и синтеза. Были выпущены опытные образцы устройств дикторозависимого распознавания изолированной речи со следующими показателями: словарь до 1000 слов, надежность распознавания 87-99 % в зависимости от размера словаря, время реакции менее 0,5 с., параметрическое представление задается 15-полосным спектроанализатором.

Таким образом, в области дикторозависимого распознавания изолированной речи существует несколько практически готовых систем распознавания слов русской речи со словарем в сотни слов. Однако как уже отмечалось выше, коммуникация, построенная на вводе отдельных слов, не обладает естественностью и скоростью взаимодействия. В результате такие системы не находят широкого применения.

Существует также и ряд более современных разработок по дикторонезависимому распознаванию русской речи, в основе которых лежат статистические модели языка. Одной из первых попыток создания дикторонезависимой системы распознавания для русского языка стала модель, разработанная исследователями компании IBM [89]. Дикторонезависимая версия системы автоматического распознавания русской речи была натренирована на 30000 высказываниях (40 русских дикторов). 3-граммная модель языка обучалась на текстах около 40 миллионов слов. Была создана система русских фонетических подгрупп и разработан набор правил для фонетического транскрибирования слов. И хотя при испытании данной модели на 8 дикторах уровень ошибки оказался не более 5%, эти исследования не получили дальнейшего развития.

Среди российских научных коллективов, которые занимаются автоматической обработкой русской речи, можно назвать СПИИРАН, ИППИ РАН, ВЦ РАН, ИСА РАН, ИПУ РАН, Московский государственный лингвистический университет, Санкт-Петербургский государственный

университет, Санкт-Петербургский электротехнический университет, Центр речевых технологий, СТЭЛ - компьютерные системы (Москва) и др.

Среди научных разработок отечественных исследователей существенных результатов добились научные группы, имеющие возможность разработки словарей и речевых корпусов большого размера. Так Институт системного анализа РАН занимается работами в области распознавания речи, которые ориентированы на развитие теоретической базы, разработку и программную реализацию методов автоматического анализа речевых сигналов в реальном масштабе времени. Предложенные решения основаны на использовании островного нейросетевого анализа речевого сигнала в корреляции с выделением устойчивых признаков и применении фонологических и других знаний о структуре речевого сигнала.

В лаборатории автоматизированных систем массового обслуживания Института проблем управления РАН более 30 лет ведутся исследования в области речевого распознавания. Главным научным и практическим направлением деятельности лаборатории является применение компьютерного распознавания слитной речи в системах обслуживания населения с возможностью использования русского и других языков. Разработаны математические модели для описания процессов в системах распознавания речи. В качестве базовой платформы для распознавания русской речи используется программное обеспечение по распознаванию речи Speech Pearl от американской компании Nuance Communication [46]. Для каждой диалоговой системы были исследованы пользовательские запросы, технологические признаки приложения и пути доступа к информации. Результатами работы лаборатории по применению распознавания речи в системах обслуживания является появление в России диалоговых систем с автоматическим голосовым интерфейсом: системы Сирена, для диспетчерской службы такси; созданы пакеты распознавания речи к службам системы Web Money [121].

В ВЦ РАН проводятся исследования и разработка методов распознавания речи, сохраняющих работоспособность в естественных условиях речевых

коммуникаций. Речь идет о том, что сейчас существует масса моделей распознавания речи, но все они созданы и проверены в лабораторных условиях, а при внедрении эти системы не обеспечивают заявленной точности. Поэтому стоит задача сохранить достаточно высокую точность распознавания в реальной ситуации, при наличии различных каналов передачи информации, шумов, неречевых акустических событий, вариабельности голосов дикторов и т.п. Общий подход состоит в использовании множественных, параллельных акустико-фонетических моделей аллофонов и неречевых акустических событий. То есть для аллофона или морфемы создаются несколько различных акустических моделей, которые совместно используются в лексической сети при декодировании речевого потока. Выбор таких множественных моделей осуществляется автоматически, путем анализа речевого корпуса данных, кластеризацией по характеру среды и голоса диктора [71].

На кафедре математической теории интеллектуальных систем и лаборатории проблем теоретической кибернетики механико-математического факультета МГУ им. М. В. Ломоносова разработан один из возможных подходов к решению проблем, препятствующих созданию промышленных систем распознавания слитной речи для русского языка. Показано, что предложенное в ней разложение общей языковой модели на две составляющие: модель, основанную на морфологии, и модель, основанную на начальных формах слов, позволяет разработчикам лучше использовать преимущества n-граммного статистического подхода. Кроме того, выделение морфологической информации в независимую модель позволяет справиться с проблемой акустической похожести различных словоформ одного и того же слова. В результате проведённых теоретических изысканий был создан пакет программ для построения различных вариантов языковых моделей для русского языка, в том числе составных моделей, основанных на категорном подходе [68].

На кафедре прикладной и экспериментальной лингвистики Московского Государственного лингвистического университета под руководством профессора Р.К. Потаповой разрабатывается модель анализа русской речи,

направленная на определение эмоционального состояния человека по речи и распознавания эмоционально окрашенной речи [102]. Значительные усилия группы исследователей направлены на создание диалоговых моделей, обработку многоязыковых лингвистических баз данных [103], а также создание теоретического фундамента науки о речи (речеведение) [32].

Следует также отметить работы отдела распознавания речевых образов ИПИИ (Донецк, Украина). Здесь были разработаны программы, которые автоматически распознают до 1000 изолированно произнесенных слов с высокой надежностью. На их основе разработана программа голосового набора математических формул в программе Equation, программа голосового управления мобильным роботом. В настоящее время отдел занимается проблемой пофонемного распознавания (фонетический стенограф) [72].

В ходе реализации совместного проекта ВНИИЭФ-СТЛ (Нижний Новгород) и Intel Corporation в 1999-2003 годах была разработана система распознавания слитной речи с большим словарем SDT (Speech Developer Toolkit) [3]. Функциональность пакета программ SDT включает: вычисление векторов признаков, построение и адаптацию акустических моделей, построение языковых моделей, быстрое декодирование речи по статистической модели или стохастической грамматике, оценку результатов декодирования. С использованием SDT были построены системы распознавания для английского и китайского языков, а также, прототип системы распознавания русской речи. В настоящее время на предприятии ВНИИЭФ-СТЛ ведутся работы по построению системы распознавания русской речи со словарем до 1 млн. слов, включая разработку компактного представления русского фонетического словаря, модифицированного алгоритма декодирования речи и статистической языковой модели для русского языка.

Центром речевых технологий разработана библиотека распознавания речевых команд VoiceCom [56]. Система обеспечивает распознавание нескольких сотен команд в дикторозависимом и нескольких десятков команд в дикторонезависимом варианте. Система может применяться для управления

технологическим оборудованием с помощью голоса; речевого запроса к базам данных; поиска ключевых слов в звуковых файлах. Также в последние годы компания начала исследование моделей русского языка, где в качестве базовых единиц распознавания взяты основы и окончания [98].

Московская компания «ИстраСофт» занимается разработками в области речевых технологий, в том числе синтезом и распознаванием речи, а также идентификацией речи по голосу. Компанией был разработан алгоритм выделения фонем из слитной речи в реальном масштабе времени. В результате работ создана программная реализация дикторонезависимого распознавания команд русской речи IstraSoft Voice Commander на основе оригинальных алгоритмов выделения звуков (фонем) в непрерывной речи [44].

Белорусской компанией «Сакрамент» разработан набор программных средств Sakrament ASR Engine [41], рассчитанный на применение в различных аппаратных системах и программных приложениях, использующих технологии распознавания речи, таких как: IVR-системы, мобильные электронные устройства, бытовая техника и т.д. Модуль Sakrament ASR Engine может быть перенесен на любую программную или аппаратную платформу, а также настроен под конфигурацию любого приложения. Качество распознавания системы зависит от размера используемых словарей, качества транскрипции, показателя связанности распознаваемых слов, уровня фонового шума, параметров используемых каналов связи и характеристик микрофонов.

Среди внедряемых в настоящее время в России систем автоматического распознавания русской речи, разработанных западными компаниями, можно отметить продукт SpeechPearl, разработанный ScanSoft и Nuance. Этот продукт является инструментарием для реализации функций распознавания речи в телефонных приложениях, который поддерживает русский язык. На этом движке построены практически все телекоммуникационные сервисы (использующие автоматическое распознавание русской речи): «Речевой портал» фирмы Светец, система Smartphone фирмы Novavox, система Telepat [52], разработанная в Институте проблем управления РАН и запущенная в

лабораторную эксплуатацию в 2004 г. Однако эти сервисы находятся пока на стадии лабораторных или тестовых прототипов. Недостатком же речевого движка SpeechPearl является ограничение по максимальному размеру распознаваемого словаря - до 5000 слов, хотя данный словарь может задаваться разработчиками сервисов и настраиваться под конкретную диалоговую модель.

Попытки западных компаний, таких как Intel Corporation или Lernout&Hauspie (Philips) создать дикторонезависимые системы диктовки для русского языка также нельзя назвать успешным. Их исследования были свернуты в связи с неудовлетворительными полученными результатами и экономическими трудностями. Среди коммерческих систем, реально дошедших до конечно пользователя, можно назвать только систему «Горыныч» [53], предложенную на Российский рынок компаниями VoiceLock и White Computers. Система имела неудовлетворительное качество распознавания (10-30% точности распознавания слов для русского языка), поскольку не учитывала особенности русской фонетики и лингвистики, а являлась лишь локализацией американской системы диктовки Dragon Naturally Speaking для русского языка. В результате разработчики перевели лишь словарь распознавания с английского на русский язык, а акустические модели и модель языка остались прежними. Поэтому, чтобы такая система работала нужно говорить по-русски, но с американским акцентом и строить грамматически правильные с точки зрения английской грамматики предложения.

Таким образом, анализ исследований по автоматическому распознаванию речи показал, что в настоящее время не существует готовых к использованию систем дикторонезависимого распознавания русской слитной речи с большим словарем, хотя решение данной задачи является очень актуальной в условиях быстрорастущего спроса на системы человеко-машинного взаимодействия.

Выводы по главе 1

1) Определены основные требования к современным системам автоматического распознавания речи: дикторонезависимость, слитный ввод речи, высокая точность распознавания и быстродействие системы, возможность работы с большими распознаваемыми словарями, робастность к различным, ухудшающим речевой сигнал, факторам.
2) Определены два интегральных критерия оценки качества систем автоматического распознавания речи: точность распознавания (слов и фраз) и время распознавания входного высказывания.
3) Проанализированы базовые подходы к распознаванию речи, основанные на скрытых Марковских моделях, искусственных нейронных сетях и динамическом программировании. В ходе анализа выявлено, что большинство современных зарубежных систем распознавания строится с использованием методов скрытого Марковского моделирования.
4) Приведено детальное описание базовой модели дикторонезависимого распознавания слитной речи, использующей вероятностные методы для акустического моделирования речи и статистического моделирования языка прикладной задачи.
5) Представлен обзор методов и технологий, применяемых для распознавания русского языка. Приводятся технические показатели основных исследовательских моделей распознавания речи с начала 40-х годов прошлого века по настоящее время. Отмечается, что сейчас не существует готовых к использованию систем дикторонезависимого распознавания русской слитной речи с большим словарем.

Глава 2. Модель обучения распознавателя русской речи с морфемным представлением языка

Любая модель распознавания речи должна функционировать в двух режимах: обучение и распознавание. Причем этап обучения является даже более сложным, так как включает в себя целый набор алгоритмов по обработке текста, статистическому анализу и вероятностному моделированию. От качества обучения, во многом, зависит качество работы модели в режиме распознавания. Во второй главе представлена предложенная архитектура модели обучения распознавателя русской речи с морфемным представлением языка прикладной области и детально описан каждый из уровней обработки, используемый в данной модели.

2.1. Особенности разработки модели распознавания русской речи

Приступая к разработке необходимо, прежде всего, проанализировать трудности, с которыми придется столкнуться в последующей разработке модели автоматического распознавания русской слитной. Основные трудности в данной задаче связаны с характерными особенностями русского языка (и других славянских языков в целом), поэтому далее приводится анализ выявленных сложностей и способов их возможного преодоления. Поскольку активнее всего ведутся исследования и разработка систем распознавания английской речи, то уместно сравнение русского и английского языков.

Русский язык относят к числу синтетических языков [12], которые характеризуются тенденцией к объединению (синтезу) лексической морфемы (или нескольких лексических морфем) и одной или нескольких грамматических морфем в рамках одной словоформы. В русском языке по сравнению с английским более сложная структура словообразования в результате чего необходимо использовать гораздо больший распознаваемый словарь, что значительно снижает как точность, так и скорость распознавания. В передовых

системах распознавания речи для английского языка (от фирм Microsoft, Nuance) используется словарь порядка 100 тыс. слов, включая распространенные имена, фамилии и названия. Для русского языка за счет наличия приставок, суффиксов и окончаний этот словарь возрастает более чем на порядок. Так грамматический словарь А.А. Зализняка [16] содержит около 100 тысяч наиболее употребительных слов русского языка, и при помощи специальной системы обозначений он позволяет построить все словоформы для выбранного слова. При развороте всех словарных статей, получается более 1,7 млн. различных словоформ. Причем данный словарь не включает в себя распространенные имена и названия, а при включении их в словарь его размер превзойдет 2 млн. словоформ. Таблица 2.1 показывает морфологические характеристики слов русского языка по основным частям речи (максимальное количество словоформ для частей речи взято из [20]). Так, например, глаголы могут формировать до нескольких сотен различных словоформ, и все их нужно учитывать при создании систем автоматического распознавания русской речи.

Таблица 2.1. Морфологические характеристики слов русского языка

Часть речи	**Грамматические категории**	**Максимальн. количество словоформ**	**Пример**
Существи-тельное	2 числа 6 падежей	32	«дерево» имеет 12 словоформ (дерево, деревом, дереве, деревьям…)
Глагол	2 залога (активный/пассивный) 2 вида (совершенный, несовершенный) 3 наклонения (изъявительное, сослагательное, повелительное) Причастия Деепричастия 3 лица, 2 числа	369	«делать» имеет 174 словоформы (делаю, делаем, делаешь, делаете, делает, делают, делал, делала, делало, делали, делаюсь, делаемся, делай, делайте, делающая, делающие, делаемого, делавши, делав, делая…)

	3 времени, 3 рода 6 падежей		
Прилага-тельное	3 сравнительных степени Краткая форма 3 рода 2 числа 6 падежей	43	«прекрасный» имеет 34 словоформ (прекрасный, прекрасная, прекрасен, прекрасней…)
Числи-тельное	Порядковые Количественные 3 рода 2 числа 6 падежей	35	«третий» имеет 29 словоформ (третий, третья, третье, третьи, третьей, третьем, третьих…)
Место-имение	3 рода 2 числа 6 падежей	51	«я» имеет 6 словоформ (я, меня, мне, мной, мною, мне)

Кроме того, большинство словоформ одного и того же слова отличаются только в окончаниях, которые произносятся обычно не так чётко как начала слов. Ошибки в окончаниях при распознавании слов приводят к тому, что происходит ошибка в распознавании всей фразы из-за несогласованности слов.

Порядок слов в предложении русского языка не задается жестко правилами грамматики и зачастую может варьироваться без потери смысла предложения, а в английском языке используются жесткие грамматические конструкции. Это затрудняет создание статистических моделей языков на основе биграмм или n-грамм, а также грамматик для русского языка и понижает их эффективность. Статистические языковые модели для русского языка не столь эффективны как для английского языка. Так в [118] показано, что n-граммные модели языка в несколько раз больше размером и их оценка неопределенности выше в 3-4 раза.

Отличия фонетического состава языков. В международном фонетическом алфавите SAMPA для русского языка принято 42 фонемы: 36 фонем согласных звуков и 6 фонем гласных звуков. В американском варианте английского языка фонетический алфавит SAMPA насчитывает 41 фонему: 24 согласных и 17 гласных (включая целый ряд дифтонгов). Очевидно, что распознавание

согласных звуков сложнее, чем гласных из-за того, что они менее стабильны, чем гласные и имеют меньшую длительность.

Для создания эффективных систем распознавания для английского языка существует несколько речевых баз данных (обычной речи, телефонной, и т.д.) в том числе и свободно доступные. Для русского языка такие базы данных пока только начинают создаваться и, как правило, являются недоступными для общего пользования.

Обозначенные выше проблемы, а также существующие экономические факторы приводят к тому, что сейчас фактически не существует действующих промышленных систем или моделей распознавания русской слитной речи.

Поэтому в диссертационной работе исследуются вопросы автоматического распознавания речи, применительно к русскому языку и речи и предложены некоторые модели, позволяющие улучшить характеристики модели распознавания русской речи.

2.2. Архитектура модели обучения с включением морфемного анализа русского языка

Одной из основных проблем автоматического распознавания речи для русского языка является обилие морфологических характеристик слов языка, что приводит к сложному механизму словообразования и наличию огромного количества словоформ. Поэтому в отличие от распознавания аналитических языков, автоматическое распознавание русской речи требует создавать словарь сверхбольшого размера, следствием чего является значительное падение точности и скорости распознавания слов и фраз.

Для решения проблемы неконтролируемого роста размера словаря при словообразовании предлагается ввести в модель распознавания дополнительный уровень представления речи – морфемный. Морфема – это наименьшая языковая единица, обладающая значением (по определению, данному американским лингвистом Л. Блумфилдом в 1933 г.) [60]. Деление морфем на части приводит только к выделению незначимых элементов языка -

фонем. За счет разделения словоформ языка на морфемы словарь распознаваемых лексических единиц может значительно сократиться, так как в процессе словообразования часто используются одни и те же морфемы.

В связи с тем, что в общепринятую архитектуру распознавания речи вводится дополнительный морфемный уровень, была модифицирована базовая модель обучения распознавателя речи, представленная в первой главе. При создании дикторонезависимой системы распознавания речи основную сложность представляет процесс обучения акустико-лексических моделей системы. Для обучения акустико-лексических единиц русской речи разработана модель, архитектура которой представлена на рисунке 2.1.

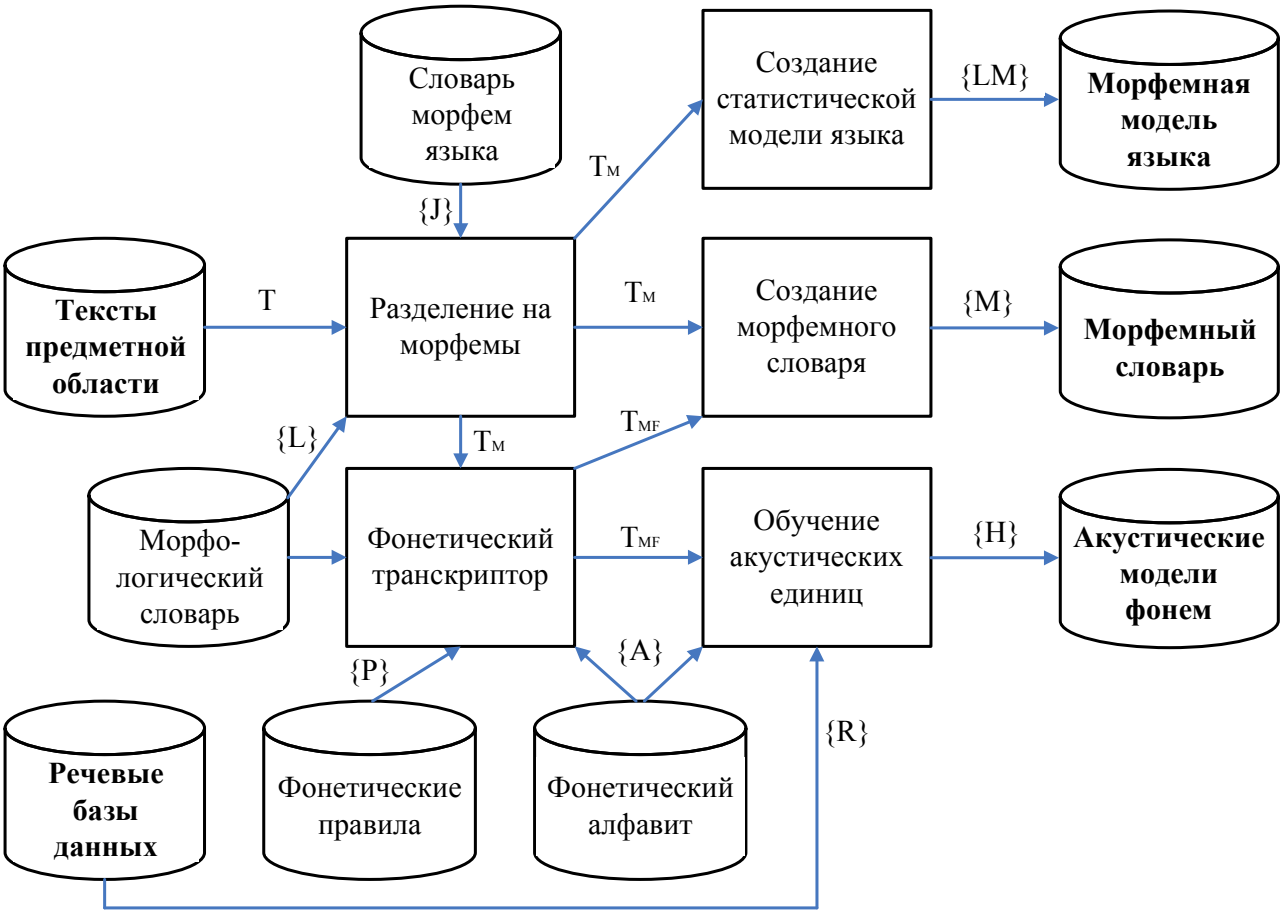

Рис. 2.1. Архитектура модели обучения распознавателя русской речи

Введем следующие математические обозначения, поясняющие предложенную архитектуру:

1) Словарь морфем языка $J = \{j_1,..., j_i,..., j_N\}$ размерностью *N,* причем $j_i = (j_{текст}, j_{тип}, j_{частьречи})$, где $j_{текст}$ - текст написания морфемы; $j_{тип} \in O = \{приставка, корень, суффикс, окончание\}$ - тип морфемы и $j_{частьречи} \in C = \{существительное, прилагательное, глагол, числительное, местоимение\}$ - часть речи, для которой морфема употребляется.
2) Морфологический словарь языка $L = \{l_1,...,l_i,...,l_K\}$ размерностью *K*, причем $l_i = (l_{словоформа}, l_{основа}, l_{частьречи}, l_{ударение})$, где $l_{словоформа}$ - текст словоформы, $l_{основа}$ - основа (часть слова без формообразующих суффиксов, окончания и постфикса) данной словоформы, $l_{частьречи} \in C$ - часть речи, к которой принадлежит словоформа и $l_{ударение}$ - место ударения в словоформе.
3) Фонетический алфавит языка $A = \{а,...,й\}$, $\alpha = |A| = 48$, количество используемых фонем равно 48 (см. далее таблицу 2.2).
4) Фонетические правила *{P}* ,более детально описаны в разделе 2.6.2.
5) Речевая база данных $R = \{r_1,...,r_i,...,r_Q\}$, причем $r_i = (\overline{F_i}, Ph_i)$, где $\overline{F_i}$ - последовательность векторов признаков звуковой записи, Ph_i - текст произнесенной фразы.
6) Текст фраз предметной области $T = Ph_1,..., Ph_i,..., Ph_F$, где F – количество предложений (фраз), состоящих из допустимых словоформ русского языка $Ph_i = w_1,..., w_j,..., w_E$, которые в свою очередь состоят из букв (графем) $w_j = b_1,...,b_h,...,b_I$, где $b_h \in B = \{а,..., я\}$ и $|B| = 33$.
7) Текст фраз предметной области с пометками границ морфем в словах и типов морфем T_M.

8) Фонетическая транскрипция текста фраз предметной области с пометками границ морфем в словах и типов морфем T_{MF} .

9) Множество акустических СММ фонем $H = \{\lambda_1,...,\lambda_\alpha\}$.

10) Морфемный словарь языка предметной области $M = \{m_1,...,m_i,...,m_z\}$, причем $1 \le Z \le N$ и $m_i = (m_{текст}, m_{транскр}, m_{тип})$, где $m_{текст}$ - текст написания морфемы, $m_{транскр}$ - фонетическая транскрипция морфемы, $m_{тип} \in \{приставка, корень, концовка\}$ - тип морфемы.

11) Морфемная модель языка $LM = \{u_1,...,u_k,...,u_s\}$, где $u_k = (m_i, m_j, P(m_i \mid m_j))$, размерностью $1 \le s \le Z^2$, причем условная вероятность $P(m_i \mid m_j) > 0$.

Работа модели обучения осуществляется за несколько этапов, первый из которых выполняется с привлечением разработчиков и экспертов в предметной области, а остальные этапы, связанные с обработкой текста, автоматизированы:

1) Подготовка баз данных предметной области. На этом этапе необходимо собрать и обработать исходный текстовый и речевой материал, который будет использован для настройки (обучения) модели распознавания речи. Необходимыми базами данных являются: корпус фраз предметной области *T* и корпус с речевым материалом *R*, записанным пользователями по части текстов предметной области. Подготовка баз данных детально описана в разделе 2.3.

2) Декомпозиция слов предметной области на морфемы. На данном этапе исходными данными является корпус текстов *T*, для которого применяется процедура декомпозиции слов на морфемы, используя для этого словарь морфем языка *J* и морфологический словарь *L.* В результате исходный текст *T* преобразуется в текст T_M с разбиением

на морфемы и для каждой морфемы проставляется ее тип. Метод декомпозиции слов детально описывается в разделе 2.4.

3) Создание статистической морфемной модели языка. Исходными данными здесь является текст с пометами морфем T_M . В результате обработки текста ищутся все пары морфем, стоящие рядом в тексте и по количеству таких событий высчитывается условная вероятность для каждой пары, составляя, таким образом, n-граммную модель языка *LM*. Создание и оценивание морфемной модели языка обсуждается в разделе 2.5.
4) Фонетическое преобразование текстов предметной области. Исходными данными являются: корпус фраз предметной области с выделением морфем T_M , морфологический словарь *L,* фонетический алфавит русского языка *A* и набор фонетических правил для транскрибирования *P*. В результате T_M преобразуется в фонетическую транскрипцию с выделением во фразах слов и морфем T_{MF} . Процесс фонетического преобразования представлен в разделе 2.6.
5) Создание морфемного словаря. Исходными данными являются: исходный текст фраз с разбиением на морфемы T_M и соответствующая ему фонетическая транскрипция T_{MF} . В ходе анализа сопоставляются два текста и выбираются уникальные морфемы, формируя морфемный словарь *M,* в котором каждой морфеме соответствует фонетическая транскрипция, причем, так как морфемы могут встречаться в различном контексте, то для морфемы возможно наличие нескольких транскрипций.
6) Обучение акустических моделей фонем. На вход данного модуля подаются: речевой корпус *R*, созданный на первом этапе, а также фонетическая транскрипция текста T_{MF} и фонетический алфавит *A*. В ходе обучения создаются скрытые Марковские модели для каждой

фонемы алфавита и переобучаются, настраиваясь на представленные речевые данные. Процесс обучения акустических моделей и формат их представления представлены в разделе 2.7.

Таким образом, в результате последовательного выполнения шагов происходит полуавтоматическое создание акустико-лексических баз данных, которые используются затем моделью распознавания слитной русской речи. При этом участие разработчика в процессе обучения требуется только на первом этапе для сбора и подготовки баз данных предметной области, все остальные шаги выполняются моделью автоматически под контролем со стороны разработчиком.

2.3. Подготовка текстовых и речевых баз данных модели обучения

Необходимой частью автоматических процедур обработки текста являются базы данных морфем и морфологический словарь, поэтому часть работы была посвящена их созданию и наполнению.

Построение базы данных морфем *J* было осуществлено на основе печатных и электронных изданий. Большая часть корневых морфем взята из открытых баз данных проекта Корнеслов [48] и словаря морфем русского языка А.И. Кузнецовой и Т.Ф. Ефремовой [13], а префиксы и флексии также из [14]. Кроме того, далее при создании ряда приложений словарь морфем постоянно пополнялся. На основе баз данных морфем можно строить любые приложения, лексика которых покрывается данным словарем. Поэтому разрабатываемые в ходе исследования базы данных можно разделить на две группы: (1) общие правила и словари всего русского языка и (2) базы данных и словари для языка конкретной предметной области (ПО). В ходе подготовки конкретного приложения также возможна модификация и дополнение общих словарей (рисунок 2.2).

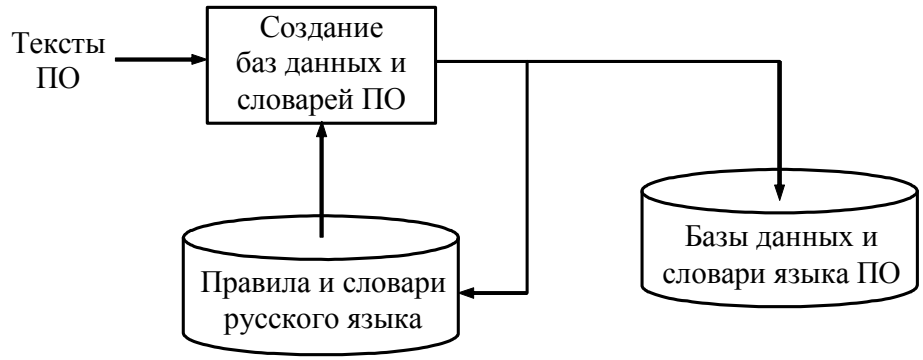

Рис. 2.2. Создание лексических баз данных для конкретной предметной области

В качестве морфологического словаря *L* в работе была использована и обработана свободно доступная морфологическая база данных, насчитывающая свыше 1.7 млн. различных словоформ [51]. Эта база данных основана на грамматическом словаре А.А. Зализняка с применением правил словообразования и расстановки ударений в словоформах (с учетом многочисленных исключений) [61].

Эти базы данных являются общими для всех приложений и при разработке новых приложений они могут только пополняться новыми данными. Остальные базы данных, требуемые для модели распознавания речи, зависят от прикладной задачи и поэтому должны создаваться и собираться разработчиками для каждого приложения отдельно.

Для создания словаря *M* модели распознавания русской речи и модели языка *LM* необходим обучающий текстовый материал языка предметной области. Для диалоговой системы такими текстами могут быть фразы или запросы пользователей, допустимые в ходе диалога, а для систем диктовки - набор предложений предметной области (желательно стенографии реальных разговоров). Однако достаточно часто получить такой материал заранее не предоставляется возможным, в таком случае для обучения используются большие массивы текстов отчетов, писем электронной почты, книг и т.д.

Последним этапом работы модели обучения распознавателя русской речи является процесс создания CMM для каждой фонемы и обучения их с

использованием речевых записей голосов различных дикторов. Для создания и обучения дикторонезависимых моделей акустических единиц речи *H* требуются речевые базы данных с записями голосов десятков или даже сотен дикторов. Для того чтобы иметь достаточное количество данных, позволяющих производить надежную оценку всех параметров модели, необходимо использовать большое множество последовательностей наблюдений.

Для сбора речевых данных в ходе работы разработан модуль, позволяющий производить запись голосов дикторов (представлен в разделе 4.1). Записи проводятся в специально оборудованной в группе речевой информатики СПИИРАН шумоизолированной комнате, где акустические условия близки к студийным. При создании этих речевых баз данных учитывается множество факторов, такие как: характеристика диктора (национальность, пол, возраст), канал передачи данных (микрофон, телефон), уровень шума. Кроме того, эти базы данных должны содержать фонетические транскрипции и разметку акустического сигнала по фонемам, словам, фразам, что является достаточно долгой рутинной ручной работой.

Чтобы избежать большого объема операций, выполняемых экспертом вручную, применяется упрощенная методика полуавтоматического создания речевых баз данных:

- Используются достаточно короткие (от 1 до 5 слов) слитно произносимые фразы из предметной области.
- Каждая фраза сохраняется в отдельном файле с применением метода автоматического удаления начальной и конечной пауз в сигнале.
- Фразы с наличием внешних шумов или артикуляторных артефактов (выдохи, шлепанье губами) отбрасываются.
- Используются только правильные фонетические транскрипции фраз, получаемые в результате анализа текстов записываемых фраз.
- Не применяется разметка границ фонем во фразах.

Таким образом, на первом этапе работы модели обучения создаются текстовые и речевые базы данных предметной области, которые используются

далее в процессе автоматического обучения модели распознавания русской слитной речи.

2.4. Декомпозиция слов предметной области на морфемы

Разделение слов на морфемы можно производить двумя путями: при помощи словарных и алгоритмических методов [80, 95]. Преимуществом алгоритмических методов является то, что они опираются лишь на анализ текста и не используют никаких дополнительных знаний, что позволяет анализировать текст на любом языке. Преимуществом словарных методов является то, что они позволяют получить правильное разбиение слов на морфемы, а не на псевдоморфемные единицы (как в алгоритмических методах), что может быть использовано далее на уровне пост-обработки гипотез распознавания фраз. Поэтому в работе использован словарный метод декомпозиции слов с использованием имеющихся для русского языка морфологических и морфемных электронных словарей.

Обычно в русском языке выделяют 6 типов морфем: префикс, корень, интерфикс, суффикс, окончание, постфикс. Были проведены эксперименты с несколькими вариантами разбиения слов на морфемы (или псевдоморфемы) и наилучшие результаты получены при разбиении слов максимально на три последовательные части: приставка (префикс), корень, концовка (псевдоокончание). Пример декомпозиции нескольких слов на морфемы показан на рисунке 2.3.

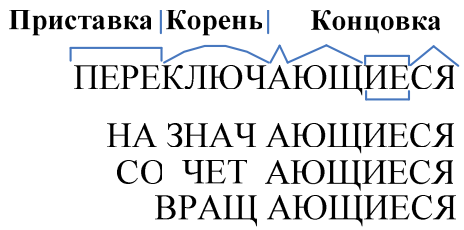

Рис. 2.3. Пример декомпозиции слов на морфемы

Такое разделение позволяет сократить количество распознаваемых лексических единиц и в то же время избежать деления слов на слишком

короткие элементы (состоящие из 1-2 букв), которые всегда тяжело распознать безошибочно.

Разбиение (декомпозиция) слова на морфемы осуществляется путем подбора морфем различных типов с учетом правил следования морфем в одном слове. Процесс преобразования слова некоторой в цепочку морфем в общем случае можно записать следующим образом:

$$D : w_i \rightarrow D(w_i) = m_1,...,m_L, \quad m_i \in J \tag{2.1}$$

где D является некоторой функцией декомпозиции слова w_i из текста T в цепочку морфем m из словаря морфем языка J.

На рисунке 2.4 показан алгоритм декомпозиции некоторого слова на морфемы, суть которого заключается в следующем:

1) Поиском в морфологическом словаре L и сравнения слова w_i со всеми значениями $l_{словоформа}$ определяется основа $l_{основа}$ анализируемого слова и его часть речи $l_{частьречи}$. Если словоформа в словаре не найдена, то слово w_i целиком считается корнем.
2) Часть слова w_i, которая следует после основы $l_{основа}$, считается концовкой слова и является самостоятельной морфемой $m_{тип} = концовка$, в случае нулевой концовки в слове, она не учитывается и не является морфемой.
3) Производя сравнение со словарем корней и приставок русского языка J, основа слова w_i разделяется (в тексте ставится разделитель «|») на две части (приставку и корень). Тут следует также учитывать, что слово может не иметь приставки и тогда основа целиком считается корнем $m_{тип} = корень$.

Таким образом, каждое слово может быть разделено максимально на 3 части: приставка (если есть), корень, концовка (если есть).

После выполнения декомпозиции всех слов из обучающего корпуса T фраз предметной области на морфемы сохраняется разметка текста на предложения, и добавляется морфемная разметка с учетом пометок о типе каждой морфемы [39]. Полученный текст T_M используется далее моделью обучения на этапе создания модели языка предметной области и фонетического преобразования.

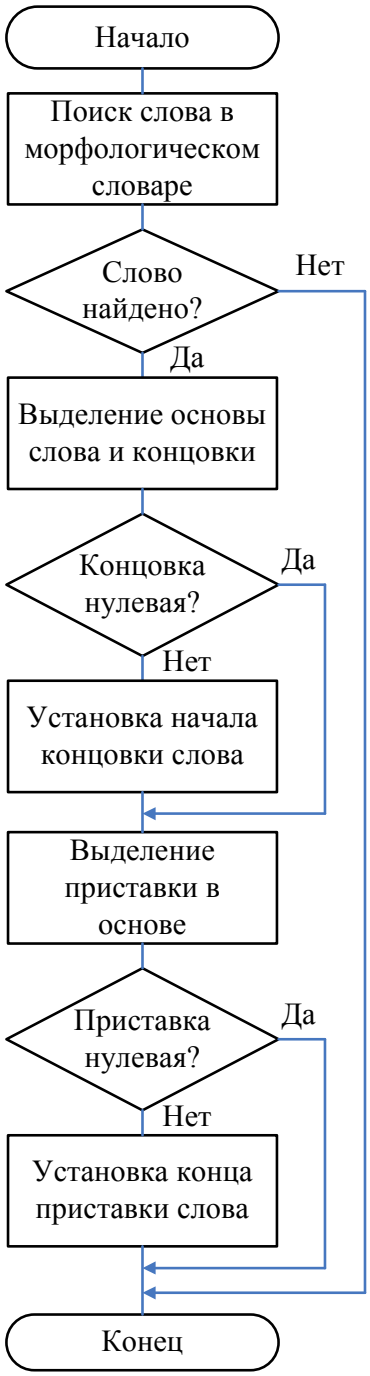

Рис. 2.4. Алгоритм декомпозиции слов языка на морфемы

2.5. Создание и оценка морфемной модели языка

В качестве модели языка применяется n-граммная статистическая модель, где лексическими единицами являются не слова, а морфемы. Ранее проводились исследования, которые показали достаточно низкую эффективность длинных цепочек слов n-граммных языковых моделей, так как в русском языке (особенно разговорном) наблюдается практически свободный порядок слов в предложении и учет истории слов во фразе недостаточно эффективен, поэтому иногда применяют модели со свободным порядком слов [69]. Использованная же в работе n-граммная морфемная модель позволяет учесть порядок стыковки морфем и правила словообразования и в то же время не задает жестких ограничений на порядок следования слов во фразе.

После декомпозиции все слов обучающего текстового корпуса T_M на морфемы, производится создание статистической морфемной модели языка. При этом формула 1.6 вычисления вероятности составления фразы из языковых единиц видоизменяется и принимает следующий вид:

$$P(Ph) = \prod_{i=1}^{V} P(m_i \mid m_{i-N+1}, m_{i-N+2}, ..., m_{i-1}), \qquad (2.2)$$

где каждое слово *w* из фразы *Ph* разделено на ряд морфем *m* посредством функции *D(w), V* – общее число морфем во фразе. Так как среднее количество применяемых морфем в слове равно двум, то используется биграммная морфемная модель языка, при этом формула 2.2 принимает следующий вид:

$$P(Ph) = \frac{1}{N(Ph)} P(m_V \mid m_{V-1}) P(m_{V-1} \mid m_{V-2}) ... P(m_2 \mid m_1), \qquad (2.3)$$

где *N(Ph)* является функцией нормализации по длине фразы.

Полный набор встреченных рядом в обучающем тексте T_M пар морфем формирует статистическую морфемную модель языка LM. При этом для каждой пары морфем оценивается сколько раз она была встречена в тексте и

применяя формулу, подобную 1.7, оценивается условная вероятность встречи данной пары морфем в тексте.

Было проведено сравнение целословной модели языка с моделями, где в качестве элементов используются языковые единицы меньше слова (морфемы, слоги). Для оценки моделей языка был использован и обработан корпус текстов классической и современной художественной литературы объемом свыше 8 млн. слов из свободно доступной электронной библиотеки М. Мошкова [42]. Для разделения слов текста на морфемы, применялся алгоритм, описанный выше в разделе 2.4.

Деление слов на слоги производилось по принципу восходящей звучности [40]. Согласно этому принципу, звуки в слоге (незаконченном) располагаются от наименее к наиболее звучному. Если звучность условно обозначить цифрами, то имеем следующую троичную классификацию: 3 – гласный звук, 2 – сонорный согласный звук ([м], [н], [л], [р], [й]), 1 – остальные (шумные) согласные звуки. Например, кни-га (1 2 3 - 1 3), и-на-че (3 - 2 3 - 1 3), по-ло-тно (1 3 - 2 3 - 1 2 3). Трудности возникают при стечении нескольких согласных подряд, для их разрешения используется несколько условий:

1) Если на границе слогов рядом оказались два шумных или два сонорных звука (кроме [й]), они относятся к последующему гласному: пу-шка, и-зба, во-лна.
2) Если в сочетании согласных первый [й], он всегда отходит к предшествующему гласному: вой-на, май-ка.
3) В сочетании согласных, первым из которых является сонорный, а вторым – шумный, сонорный может отходить к предшествующему гласному: кон-спект, Вол-га.

Любой язык характеризуется словарным составом и структурной организацией слов [1]. Статистические модели языка количественно могут оцениваться рядом параметров [88, 118]: количество лексических единиц в модели (размер словаря), коэффициент неопределённости, процент «непокрытия» слов (out-of-vocabulary) в тестовом текстовом материале и т.д.

Были проведены эксперименты по оценке моделей языка, результаты которых представлены на рисунке 2.5, отражающем график количества различных лексических единиц в зависимости от объема текста, а также на рисунке 2.6, отражающем процент непокрытых слов в тексте. В ходе экспериментов в качестве тестового текста для каждой модели использовался полный текст романа М.А. Булгакова «Мастер и Маргарита», а в качестве обучающих текстов применялся электронный корпус, состоящий из нескольких десятков текстов произведений классической художественной литературы (не включая тестовый текст) [42].

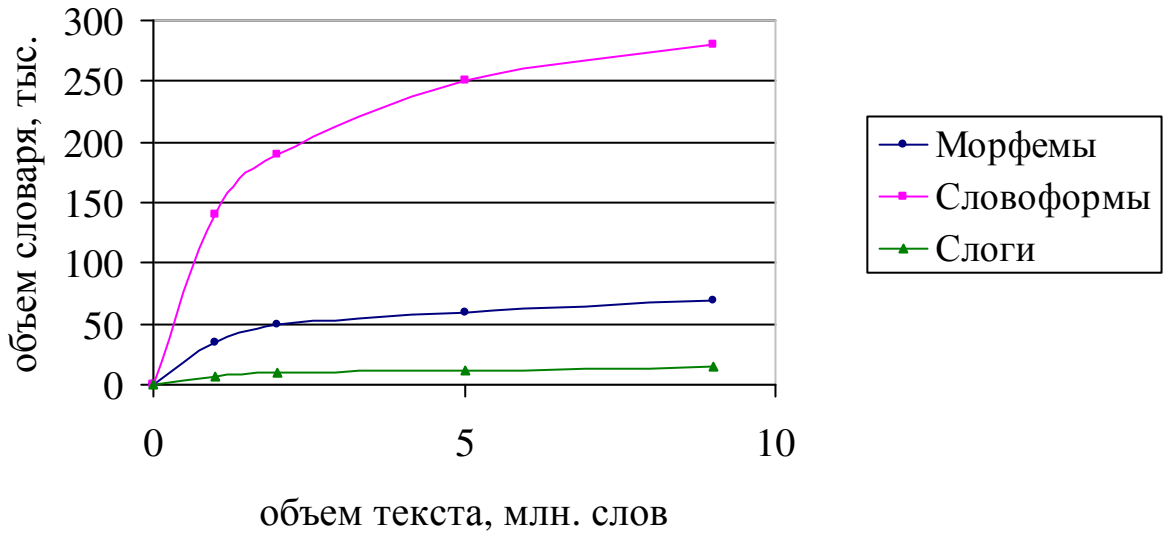

Рис. 2.5. Размер словаря языковых единиц в зависимости от объема обучающего текстового корпуса

На рисунке 2.5 представлены графики количества различных словоформ, морфем и слогов, которые встречаются в обучающем текстовом корпусе и, таким образом, потенциально составляют словарь распознавания. Так, например, при анализе корпуса текстов из 2 млн. слов можно выделить свыше 190 тыс. различных словоформ, 50 тыс. различных морфем (большая часть из них - корни) и до 7 тыс. слогов. Таким образом, размер распознаваемого словаря при использовании словоформ в 4 раза больше, чем для морфемного словаря и в 27 раз больше, чем для слогового словаря. При увеличении размера

обучающего корпуса это соотношение становится еще больше. Характеристика модели языка, представленная на рисунке 2.6, особенно важна при разработке стенографических систем, так как отражает процент слов, которые встречаются в тестовых записях, но не встречаются в обучающих и, таким образом, не могут быть правильно распознаны. Так, графики показывают, что если использовать словарь распознавания в 50 тыс. словоформ, то он не покрывает около 20% слов тестового материала, морфемный же словарь не позволяет покрыть менее 1% слов текста, а слоговый словарь лишь 0,1%.

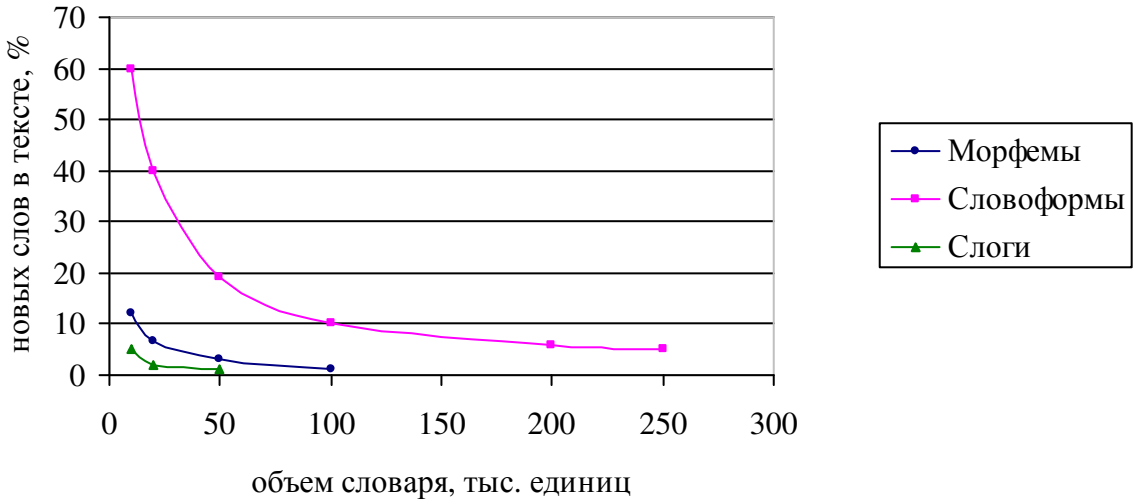

Рис. 2.6. Процент «непокрытых» слов тестового текста в зависимости от размера распознаваемого словаря

Таким образом, приведенные графики показывают преимущество слоговой и морфемной моделей языка перед целословной моделью. Из приведенного анализа можно сделать вывод, что чем меньше размер используемых лексических единиц, тем меньше их встречается в языке и тем более полно они покрывают пространство слов языка. Так, например, цепочки морфем могут образовывать правильные словоформы, которых даже не было в словаре распознаваемых слов. В четвертой главе будут представлены результаты использования целословной, морфемной и слоговой моделей

распознавания русской речи, которые показывают эффективность двух первых моделей и неэффективность использования третьей.

2.6. Фонетическое транскрибирование обучающих текстов

Моделирование речи для систем распознавания с большим словарем должно вестись по фонемам (звукам), так как практически невозможно собрать акустический материал для надежного обучения дикторонезависимых целословных акустических моделей. Любой речевой поток представляет собой непрерывную линейную последовательность звуков, при этом звук является минимальной незначимой единицей речевого потока и выступает в речи нерасчлененно, как единое целое и из него уже нельзя выделить отдельные более мелкие элементы произношения (артикуляции) [40].

2.6.1. Выбор фонетического алфавита

В русском языке слова обычно произносятся не так как они пишутся, поэтому необходим фонетический транскриптор, производящий преобразование «буква->фонема» [59]. Для передачи звучащей речи на письме используется особая запись, отличная от орфографической - фонетическая транскрипция. Фонетическая транскрипция - это запись звуков слов языка с учетом их позиционных изменений в речевом потоке.

Первым этапом в разработке фонетического транскриптора является выбор фонетического алфавита, который содержит символы транскрипции. Особенности тех или иных звуков могут фиксироваться посредством специальных символов транскрипции. Разнообразие звуков речи чрезвычайно велико, поэтому в каждой из систем транскрипции, кроме основных символов, имеются многочисленные дополнительные знаки. Наиболее распространены системы, использующие в качестве фонетических символов знаки латиницы с добавлением знаков из других графических систем, например, система МФА (Международной фонетической ассоциации) [87] или фонетическая система Л.В. Щербы [73]. В этих системах имеются символы для обозначения

согласных и гласных и их дополнительных артикуляционных свойств (палатализованность, веляризованность, придыхательность, огублённость, назализованность, отодвинутость назад, продвинутость вперёд), для обозначения степеней силового ударения, тона и характера музыкального ударения и т.д.

В последние годы также была предложена система символов фонетической транскрипции SAMPA (Speech Assessment Methods Phonetic Alphabet) для многих языков [54]. В транскрипции SAMPA принято использовать только те символы, которые имеются на клавиатуре персонального компьютера. С их помощью оказывается возможным передавать фонетическую информацию по межкомпьютерной связи.

В диссертационной работе в качестве фонетического алфавита предложен и опробован модифицированный вариант международного фонетического алфавита SAMPA. Предложенный фонетический алфавит, а также примеры слов представлены в таблице 2.2. В данном наборе используются также специальные символы транскрипции: «!» - обозначает ударный вариант гласного звука; «'» - обозначает мягкий вариант согласного звука.

Таблица 2.2. Фонетический алфавит для распознавателя русской речи

Фонема	**Слово**	**Транскрипция**	**Фонема**	**Слово**	**Транскрипция**
а	пара	п а! р **а**	**ц**	цепь	**ц** э! п'
а!	пара	п **а!** р а	**ч**	чай	**ч** а! й
и	мели	м' е! л' **и**	**ф**	фарс	**ф** а! р с
и!	мир	м' **и!** р	**ф'**	физика	**ф'** и! з' и к а
е	дерево	д' е! р' **е** в а	**в**	ваза	**в** а! з а
е!	дерево	д' **е!** р' е в а	**в'**	виза	**в'** и! з а
ы	дыры	д ы! р **ы**	**с**	сын	**с** ы! н
ы!	дыры	д **ы! р** ы	**с'**	сено	**с'** е! н а
у	тулуп	т **у** л у! п	**з**	запах	**з** а! п а х
у!	тулуп	т у л **у!** п	**з'**	корзина	к а р **з'** и! н

о!	город	г **о!** р а т	**ш**	шар	**ш** а р
э!	цепь	ц **э!** п'	**щ**	щука	**щ** у к а
п	пыль	**п** ы! л'	**ж**	жир	**ж** ы! р
п'	пить	**п'** и! т'	**х**	хлеб	**х** л' е! п
б	быть	**б** ы! т'	**х'**	хитрый	**х'** и! т р ы й
б'	бить	**б'** и! т'	**м**	май	**м** а! й
т	тост	**т** о! с т	**м'**	мята	**м'** а т а
т'	тень	**т'** е! н'	**н**	найти	**н** а й т' и!
д	дым	**д** ы! м	**н'**	нить	**н'** и! т'
д'	день	**д'** е! н'	**л**	луч	**л** у! ч
к	кот	**к** о! т	**л'**	любовь	**л'** у б о! ф'
к'	кит	**к'** и! т	**р**	краб	к **р** а! п
г	гусь	**г** у! с'	**р'**	резать	**р'** е! з а т'
г'	гибкий	**г'** и! п к' и й	**й (j)**	июль	и **й** у! л'

В предложенном фонетическом алфавите используется 48 фонем: 12 - для гласных звуков (с учетом ударных вариантов) и 36 - для согласных (с учетом твердости и мягкости звуков). Модификация международного алфавита заключается в добавлении к обычным вариантам гласных звуков ударных вариантов некоторых гласных звуков. Так как ударные и безударные гласные имеют значительные отличия в спектральных и временных характеристиках, то такое разделение позволяет улучшить точность описания и акустико-фонетического моделирования речи.

2.6.2. Фонетическое транскрибирование текста

Модуль фонетического транскрибирования осуществляет преобразование текста предметной области с пометами морфем T_M в его фонетическое представление T_{MF}. На рисунке 2.7 показан алгоритм фонетического транскрибирования некоторого слова из текста T_M.

В ходе работы алгоритма слово w_i из текста ищется в морфологическом словаре L и в случае нахождения совпадения со словоформой $l_{словоформа}$ определяется место ударения в слове $l_{ударение}$ и помечается знаком «!». В случае отсутствия такого слова в словаре и наличии более одной гласной в слове, невозможно создать для него автоматически правильную фонетическую транскрипцию из-за возможной неоднозначности в месте ударения.

После нахождения места постановки ударения в слове w_i алгоритм применяет к слову правила фонетического преобразования P, при этом возможны следующие позиционные изменения классов звуков: изменения гласных в положении под ударением; изменения гласных в предударных слогах; изменения гласных в заударных слогах; позиционные изменения согласных звуков [40].

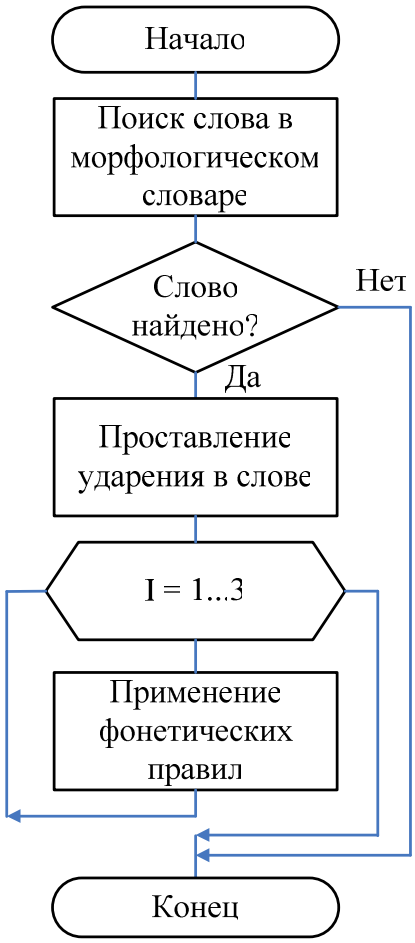

Рис. 2.7. Алгоритм фонетического транскрибирования слова

Применение фонетических правил выполняется за 3 цикла, в процессе каждого из которых над каждым транскрибируемым словом текста последовательно применяются следующие фонетические правила:

1) Согласные звуки перед буквами *и, е, ё, ю, я* становятся мягкими (приобретают знак «'»).
2) Буквы *ё, ю, я* заменяются на сочетания звуков [йо], [йу], [йа] соответственно, если они находятся в начале слова или после гласного звука, в противном случае они заменяются на гласные звуки [о], [у], [а].
3) Согласный звук перед буквой *ь* становится мягким, а сам мягкий знак убирается.
4) Парные звонкие согласные в конце слова становятся глухими.
5) Оглушаются согласные перед глухими шумными и озвончаются согласные перед звонкими шумными согласными.
6) Смягчаются согласные перед мягкими зубными согласными звуками [т'], [д'], [с'], [з'].
7) Изменения гласных звуков под ударением и безударных гласных в предударном слоге [40].
8) Изменения гласных звуков в заударном слоге [40].
9) Изменения двухбуквенных сочетаний согласных (включая дублирование согласных).
10) Из текста удаляется твердый знак ъ.
11) В конце слов перед окончанием «о» звук [г] заменяется на звук [в] (например, в слове «белого»). Для выполнения такой замены используется информация о разбиении слова на морфемы.
12) Изменения многобуквенных сочетаний согласных (например, слово «солнце» представляется фонетической транскрипцией [со!нце]).

Применяя приведенные правила, текстовый морфемный корпус T_M преобразуется (отображается) в фонемное представление T_{MF}, сохраняя пометки границ морфем:

$$P : T_M \rightarrow T_{MF} \qquad (2.4)$$

В таблице 2.3 представлен пример автоматического фонетического транскрибирования фрагмента текста художественного произведения «Мастер и Маргарита» М.А. Булгакова.

Таблица 2.3. Пример фонетического транскрибирования фрагмента текста

Исходный текст	Фонетическая транскрипция
Никогда не разговаривайте с неизвестными. Однажды весною, в час небывало жаркого заката, в Москве, на Патриарших прудах, появились два гражданина. Первый из них, одетый в летнюю серенькую пару, был маленького роста, упитан, лыс, свою приличную шляпу пирожком нес в руке, а на хорошо выбритом лице его помещались сверхъестественных размеров очки в черной роговой оправе. Второй - плечистый, рыжеватый, вихрастый молодой человек в заломленной на затылок клетчатой кепке - был в ковбойке, жеваных белых брюках и в черных тапочках.	н'и/кагда! н'и раз/гава!р'/ивайт'е с н'и/изв'е!с/ным'и адна!жд/ы в'исн/о!йу ф ча!с жа!р/кава за/ка!т/а ф маскв'/е на патр'иа!рш/ых пруд/а!х па/йив'/и!л'ис' два! граждан'и!н/а п'е!рв/ый ис н'и!х ад'е!т/ый ф л'е!т/н'уйу с'е!р'/ен'куйу па!р/у бы!л ма!л'/ен'кава ро!ст/а у/п'и!т/ан лы!с сва/йу! пр'ил'и!чн/уйу шл'а!п/у п'ира/шко!м н'ос ф рук'/е! а! на хараш/о! вы!/бр'ит/ам л'иц/э! йиво! па/м'ищ/а!л'ис' св'е!рх/йис'т'е!ств'/еных раз/м'е!р/аф ачк'/и! ф чо!р/най раг/аво!й а/пра!в'/е фтар/о!й пл'ич/ы!стый рыж/ыва!тый в'ихр/а!стый малад/о!й чилав'е!к ф за/ло!м/л'енай на за/ты!л/ак кл'е!ч/атай к'е!пк'/е бы!л ф кавбо!й/к'е *жеваных* б'е!л/ых бр'у!к/ах и! ф чо!р/ных та!пачк/ах

Из данного примера видно, что существуют некоторые ошибки при фонетическом транскрибировании. Так слово «жеваных» не было найдено в словаре словоформ с ударениями и поэтому транскрипция для него отсутствует

по причине невозможности проставления места ударения для данного слова. Вторая наиболее частая причина ошибок - неверное проставление ударения в слове, частично решается за счет синтаксического разбора предложений.

Применение разработанного модуля фонетического транскрибирования позволяет получить правильную фонетическую транскрипцию примерно для 95% слов. Ошибки встречаются для слов-исключений, омографов и некоторых слов, заимствованных из других языков, однако этими ошибками можно пренебречь или исправить ошибки вручную (при небольшом объеме обучающих текстов).

Таким образом, создается фонетическая транскрипция текста предметной области, что позволяет далее использовать эту информацию для создания морфемного распознаваемого словаря (с фонетическим описанием), а также обучения СММ фонетико-акустических единиц речи и создания морфемного словаря.

2.7. Создание и обучение моделей акустико-фонетических единиц речи

Для получения СММ фонем русского языка из фонетического алфавита A сначала требуется создать начальные СММ для каждой фонемы и проинициализировать их параметры. При этом начальные значения параметров СММ могут получаться несколькими способами, включая ручную сегментацию последовательностей наблюдений на состояния с усреднением числа наблюдений в состояниях, сегментацию наблюдений по методу максимального правдоподобия с усреднением, сегментацию с использованием метода k-средних. Начальная оценка моделей может выбираться произвольным образом или же на основе любой имеющейся модели, соответствующей этим данным. В работе применяется метод k-средних с равномерным распределением вероятностных параметров при инициализации СММ [119].

Для процесса обучения СММ фонем (рисунок 2.8) входными данными являются: начальные СММ всех фонем с равномерными вероятностными оценками параметров моделей, речевые записи из базы данных R и

соответствующие им тексты записанных фраз с фонетическими транскрипциями из текста T_{MF} .

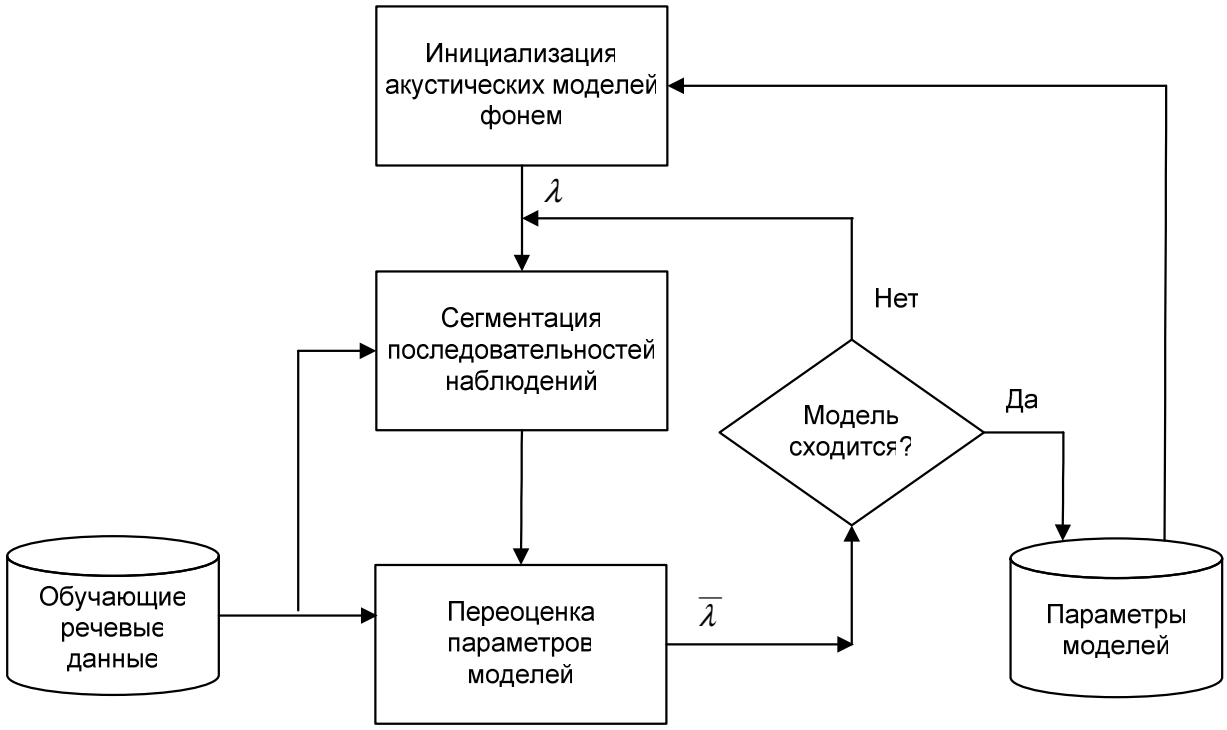

Рис. 2.8. Процесс обучения параметров СММ фонем языка

Процесс обучения параметров СММ фонемы $\lambda_i = (N, A, B, n)$ из набора моделей *H* осуществляется следующим образом:

- Переоценка параметров СММ фонем на записях обучающих речевых баз данных с использованием алгоритма Витерби (описан в главе 1).
- Когда простая переоценка параметров СММ не дает улучшения, происходит увеличение числа компонент смесей нормальных распределений в состояниях моделей с одновременным обучением на речевых данных. При этом распределение векторов наблюдений описывается смесью нормальных распределений: $b_i(O) = \sum_{m=1}^{G} C_{jm}\, \vartheta[O, \mu_{jm}, U_{jm}]$, где O – моделируемая последовательность векторов, C_{jm} - коэффициент смеси для *m*-ой

компоненты в состоянии *j*, Гауссова плотность вероятности:

$$\vartheta[O,\mu,U] = \frac{1}{\sqrt{(2\pi)^n |U|}} e^{-\frac{1}{2}(O-\mu)^T U^{-1}(O-\mu)}, \quad (2.5)$$

где μ - математическое ожидание, U диагональная ковариационная матрица, *n* – размерность векторов наблюдений *O*.

Критерием оптимизации параметров СММ является критерий максимального правдоподобия модели λ на наборе обучающих данных *O*:

$$\lambda_{оптим} = \arg\max P(O \mid \lambda) \quad (2.6)$$

Процедура переоценки параметров выполняется итеративно, используя на каждом новом шаге значения параметров модели, полученные на предыдущем шаге. Обновленная модель $\overline{\lambda}$ получается на основе новых параметров модели, а переоценка всех параметров этой модели выполняется с помощью формальной процедуры повторного оценивания. Затем результирующая модель $\overline{\lambda}$ сравнивается с предыдущей моделью λ посредством вычисления меры отклонения (разности вероятностей наблюдений для моделей), которая отражает статистическое сходство этих моделей. Если эта мера отклонения превышает порог, то старая модель заменяется новой и полностью повторяется цикл обучения. Если же мера отклонения не превышает данного порога, то полагается, что модель сходится, и сохраняются параметры последней модели.

Таким образом, в ходе обучения последовательно изменяются параметры СММ, при этом вероятность появления обучающих последовательностей наблюдений увеличивается. Выполнение процедур продолжается до тех пор, пока не будет достигнута некоторая предельная точка (по критерию максимума правдоподобия СММ). На выходе модуля обучения СММ акустических единиц создаются модели фонем, моделируя, таким образом, голос «среднего» диктора (или нескольких таких кластеров) и возможные отклонения по каждому из параметров.

Выводы по главе 2

1) Проведен анализ основных проблем, связанных с распознаванием русской речи. Выявлена проблема большого количества словоформ в языке из-за богатых механизмов словообразования, что значительно увеличивает размер распознаваемого словаря, а также проблема отсутствия жестких грамматических правил построения предложений, допускающая практически свободный порядок слов во фразе.
2) Для решения проблемы значительного роста размера словаря при формировании различных словоформ слова предлагается проводить распознавание речи не по словам, а по морфемам. Предложена модель обучения распознавателя русской речи с большим словарем с включением морфемного уровня анализа русского языка.
3) Разработан метод декомпозиции слов языка предметной области на морфемы (псевдоморфемы), что позволяет значительно сократить количество языковых единиц в процессе распознавания и ускорить, таким образом, автоматическую обработку речи.
4) Произведено построение целословной, морфемной и слоговой статистических моделей языка и сравнение по критериям размера распознаваемого словаря и проценту непокрытых слов в тестовом тексте и выявлено преимущество морфемной и слоговой моделей.
5) Предложен фонетический алфавит для задачи автоматического распознавания русской речи, включающий набор из 48 фонем и отличающийся учетом ударных и безударных вариантов гласных звуков в речи. Разработана процедура фонетического транскрибирования текста, а также создания морфемного словаря модели распознавания русской речи.

Глава 3. Модель распознавания русской речи с морфемным уровнем обработки

В ходе сложного многоуровневого процесса распознавания речи звуковой сигнал, получаемый от микрофона, преобразуется в последовательность слов, составляющих гипотезу распознавания фразы. Первым этапом функционирования любой модели распознавания речи является процесс обучения, в ходе которого настраиваются акустические модели и подготавливаются лексические базы данных. Процесс обучения модели распознавания речи был представлен в главе 2, данная глава описывает второй этап функционирования - автоматическое распознавание русской речи.

В отличие от других моделей в предложенной модели распознавания русской слитной речи распознаваемыми языковыми единицами вместо слов являются морфемы (рисунок 3.1). На рисунке 3.1 выделены блоки, в которых произошли изменения по сравнению с базовой моделью распознавания слитной речи, представленной выше на рисунке 1.2.

Подготовленные базы данных конкретной предметной области, а именно: словарь морфем языка предметной области (ПО) с их фонетическим представлением, морфемная модель языка, фонемный алфавит и набор обученных акустических моделей фонем используются в ходе процесса автоматического распознавания русской речи. Однако если при подготовке баз данных производился анализ фраз, разбивая их на морфемы и фонемы, то в ходе распознавания речи осуществляется обратный процесс: из наиболее вероятных цепочек фонем последовательно синтезируются морфемы, слова и фразы (рисунок 3.1).

Любой звуковой сигнал, поступающий с микрофона или считываемый из файла, сначала проходит этап начальной обработки сигнала, где отрезаются начальные и конечные паузы, а оставшийся речевой сигнал преобразуется в последовательность векторов признаков, которая поступает в модуль

распознавания фонем, где используются методы скрытого Марковского моделирования, описанные в первой главе.

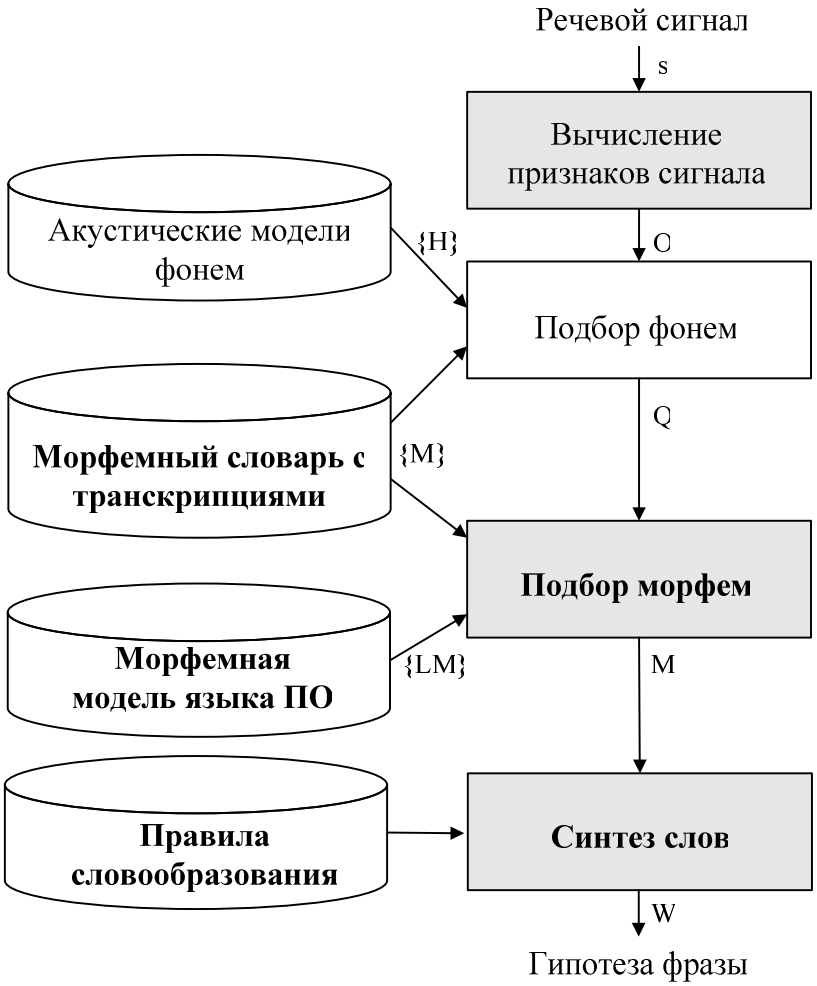

Рис. 3.1. Архитектура модели дикторонезависимого распознавания русской слитной речи с морфемным уровнем обработки

В ходе диссертационного исследования автором были проработаны все основные уровни обработки сигнала. На этапе начальной обработки звукового сигнала предложена модификация метода определения границ речи из фонового шума, основанного на анализе функции энтропии спектра сигнала, а также метод спектрально-разностного представления речи. Эти методы будут представлены далее в разделах 3.1 и 3.2, соответственно. В заключительном разделе третьей главы представлен метод распознавания слов слитной русской речи, использующий морфемный уровень представления языка и речи.

3.1. Выделение речи в звуковом сигнале методом анализа спектральной энтропии

Традиционно в системах распознавания речи для определения границ речи используются методы, основанные на вычислении кратковременной энергии сигнала или спектральной энергии (например, Voice Activity Detector [81]). Кроме того, дополнительно применяются методы, использующие количество нуль-пересечений сигнала и информацию о длительности речевых фрагментов [106]. Однако эти алгоритмы становятся менее надежными в условиях нестационарного шума, а также при возникновении различных звуковых артефактов (придыхание, чмокание и т.п.). Также существуют алгоритмы, основанные на адаптивных пороговых значениях, но при возникновении звуковых артефактов, а также относительно высоком уровне шума или незначительном уровне полезного сигнала они также становятся не устойчивыми. Поэтому в ходе исследований была поставлена задача разработать эффективный метод для определения границ речи, который позволил бы устойчиво выделять речь при наличии нестационарного шума. При этом к методу определения границ речи предъявляются следующие основные требования:

- Обеспечение минимальной вероятности ложного срабатывания при воздействии только шума с высоким уровнем.
- Высокая вероятность правильного выделения речи даже в условиях сильного шума.
- Высокое быстродействие для исключения задержек включения и выключения распознавателя речи.

3.1.1. Математическая основа метода

Предложенный метод основан на вычислении информационной энтропии спектра сигнала. Для определения границ речи используется свойство отличия значений энтропии для речевых сегментов сигнала и для сегментов фонового шума. Отличительная черта данного подхода состоит в том, что этот

показатель является мало чувствительным к изменениям амплитуды сигнала. Предложенный метод является развитием существующих идей [108, 117] и добавляет новый уровень при анализе звукового сигнала. Рисунок 3.2 иллюстрирует процесс определения границ речи на основе спектральной энтропии.

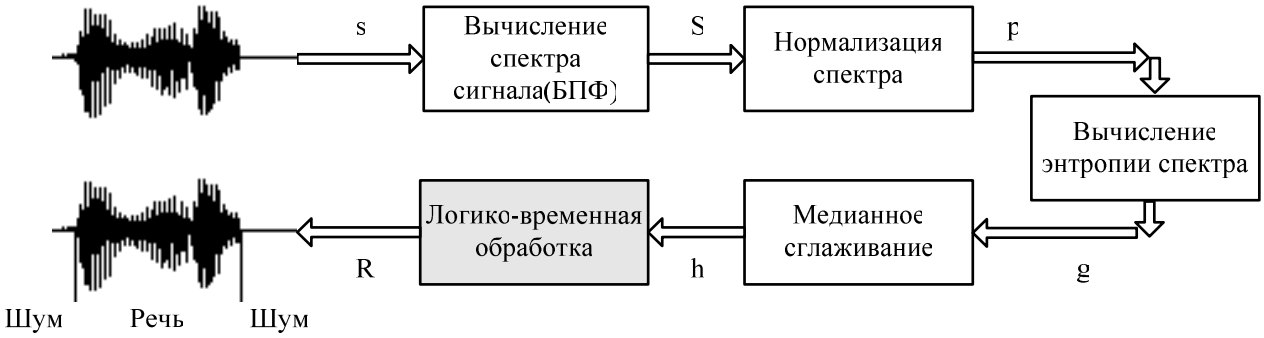

Рис. 3.2. Определение границ речи в сигнале на основе анализа функции энтропии спектра сигнала

Поступающий с микрофона сигнал оцифровывается с частотой дискретизации 16 КГц и делится на короткие сегменты сигнала *s(m)* по 16 мс, содержащие по 256 цифровых отсчетов. При этом перекрытие соседних сегментов составляет 70 отсчетов (более 25%), чтобы устранить краевые эффекты. Далее рассчитываются мгновенные спектры мощности сигнала как квадраты амплитуд преобразования Фурье для каждого сегмента *i*:

$$S_i(k) = \left| \sum_{m=0}^{M-1} s_i(m) e^{-j\frac{\pi}{M}mk} \right|^2, \quad 0 \le k \le M \tag{3.1}$$

где M – размер сегмента сигнала.

Затем производится нормализация вычисленного спектра по всем частотным компонентам:

$$p_i = \frac{S(f_i)}{\sum_{k=1}^{N} S(f_k)} \tag{3.2}$$

Полученная функция представляет собой функцию плотности вероятности спектра. Возможные значения плотности вероятности спектра ограничиваются как сверху, так и снизу, что позволяет исключить шумы, сосредоточенные в узкой частотной области, а также шумы, имеющие приблизительно одинаковое распределение частотных компонент по всему спектру (например, белый шум).

$$p_i = \begin{cases} 0, p_i > \delta_1 \\ 0, p_i < \delta_2 \\ p_i, иначе \end{cases}, \tag{3.3}$$

где δ_1 и δ_2 - верхняя и нижняя границы плотности вероятности спектра сегмента. В ходе экспериментов значения δ_1 и δ_2 были выбраны 0.3 и 0.01, соответственно при возможных значениях p от 0 до 1. Дополнительно могут использоваться методы очистки сигнала от шума (например, адаптивный фильтр Кальмана или методы спектрального вычитания) [75, 82].

Ключевым этапом метода обработки сегмента речи является вычисление спектральной энтропии (как меры неопределенности или беспорядка в некотором распределении) полученного нормированного спектра. Для оценки информативности сегмента, путем вычисления информационной энтропии, используется следующая формула [70]:

$$H = -\sum_{k=1}^{N} p_k \log p_k \tag{3.4}$$

Чтобы избежать случайных выбросов значений спектральной энтропии необходимо произвести сглаживание функции спектральной энтропии сигнала g для получения сглаженной функции h. Предлагается использовать медианное сглаживание последовательности полученных значений спектральной энтропии. В отличие от других методов сглаживания (например, метода скользящих средних), данный метод является более устойчивым к отдельным выбросам и случайным искажениям данных. В основе метода лежит вычисление скользящей медианы [5]. Для того чтобы найти значение скользящей медианы в точке t, вычисляется медиана значений ряда во

временном интервале *[t-q, t+q]*. Медиана ряда во временном интервале определяется как центральный член последовательности значений ряда, входящих в этот временной интервал, упорядоченной по возрастанию. В ходе экспериментов наилучшие результаты показал метод медианного сглаживания в окне размером 5 сегментов. Однако, если момент времени *t* отстоит от начала или конца ряда менее чем на *q* точек, вычисление становится невозможным. Поэтому здесь для устранения таких краевых эффектов вычисляется значение скользящей медианы для меньшего, но максимально возможного окна.

На следующем уровне обработки вычисляется адаптивный порог, который служит для выделения краевых точек (начала и конца) гипотезы фрагмента речи:

$$r = \left(\frac{\max(h) - \min(h)}{2} + \min(h) \right) * \mu \qquad (3.5)$$

где μ - коэффициент зашумленности, который подбирается экспериментальным путем в зависимости от характера и интенсивности шума. Опытным путем установлено, что данный коэффициент может принимать значения от 0,8 до 1,1 в зависимости от зашумленности сигнала. На основе адаптивного порога *r* выбираются акустические сегменты анализируемого сигнала, которые являются речью.

На последнем этапе анализа сигнала применяется логико-временная обработка (рисунок 3.3) вычисленной функции энтропии спектра сигнала *h*, учитывающая допустимые на практике длительности речевых и неречевых фрагментов, определенных с применением адаптивного порога *r* по формуле 3.5. Эта обработка необходима, так как во многих случаях из-за возникновения таких звуковых артефактов как чмоканье или щелканье безречевые участки сигнала ошибочно принимаются за речь, и наоборот, некоторые участки, содержащие речь, отбрасываются из-за специфических акустических характеристик. Применяя адаптивный порог к функции *h,* мы можем определить чередующиеся речевые и безречевые участки на функции h и применить для обработки два критерия:

1) R - минимальная длительность речевого участка.
2) S - максимальная длительность безречевого участка между соседними речевыми фрагментами.

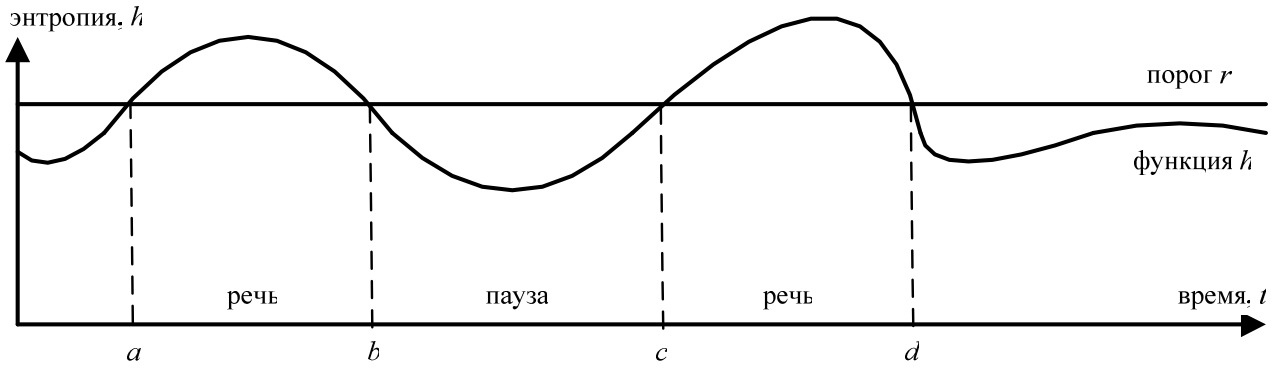

Рис. 3.3. Логико-временная обработка функции энтропии спектра сигнала

Учитывая, что человек не может производить очень короткие речевые фрагменты, а также то, что в речи всегда присутствуют определенные паузы (например, смычки перед взрывными согласными) эмпирически определены значения минимальной длительности речевого участка R и максимальной длительности безречевого участка S (15 и 20 сегментов соответственно). Анализ обнаруженных речевых фрагментов и неречевого фрагмента между ними происходит по следующей формуле:

$$речь = \begin{cases} ad, (ab \geq R) \wedge (cd \geq R) \wedge (bc \leq S) \\ ab, (ab > cd) \wedge (ab \geq R) \\ cd, (ab < cd) \wedge (cd > R) \\ \varnothing, иначе \end{cases} \qquad (3.6)$$

Данное правило итеративно применяется ко всем соседним размеченным фрагментам анализируемой функции спектральной энтропии.

Таким образом, если речевые участки в некотором сигнале имеют длительности не менее R сегментов, а безречевой участок между ними – не более S сегментов, то все данные фрагменты участки объединяются в один речевой фрагмент, который и будет являться результатом работы метода.

3.1.2. Экспериментальная проверка метода

В ходе диссертационной работы был проведен ряд экспериментов по выделению раздельно произнесенных слов, а также слитно произнесенных фраз из звукового сигнала. Для проверки работоспособности метода были сгенерированы и внедрены в сигнал следующие виды искусственных шумов (рисунок 3.4):

1) Шум с узкой полосой частот (2700 - 3300 Гц). Данный шум можно приближенно считать монотонным сигналом с частотой 3000 Гц.
2) Белый шум. Данный шум имеет спектр с приблизительно постоянной спектральной плотностью в полосе частот от 0 до 8000 Гц.
3) Коричневый шум. Спектральная плотность уменьшается на 6 дБ с каждой последующей октавой (т.е. спектральная плотность обратно пропорциональна квадрату частоты).
4) Розовый шум. Спектр такого шума имеет спектральную плотность, уменьшающуюся на 3 дБ с каждой последующей октавой (спектральная плотность обратно пропорциональна частоте).
5) Усиленный в 10 раз акустический фон, записанный в помещении, в котором производились эксперименты.

Рисунки 3.4 и 3.5 иллюстрируют первый пример работы метода по выделению произношения слова «аттракцион» из сигнала, в котором присутствуют все вышеперечисленные виды шумов с амплитудами большими или равными произнесенной речи. Рисунок 3.5 показывает график функции спектральной энтропии сигнала, а горизонтальная линия на нем – значение адаптивного порога (вычисляемого по формуле 3.5), серым цветом выделен участок сигнала, в котором предложенный метод обнаружил речь. Очевидно, что речь в данном примере была выделена точно. Также можно заметить, что со всеми видами шумов алгоритм справился, не перепутав их с речью.

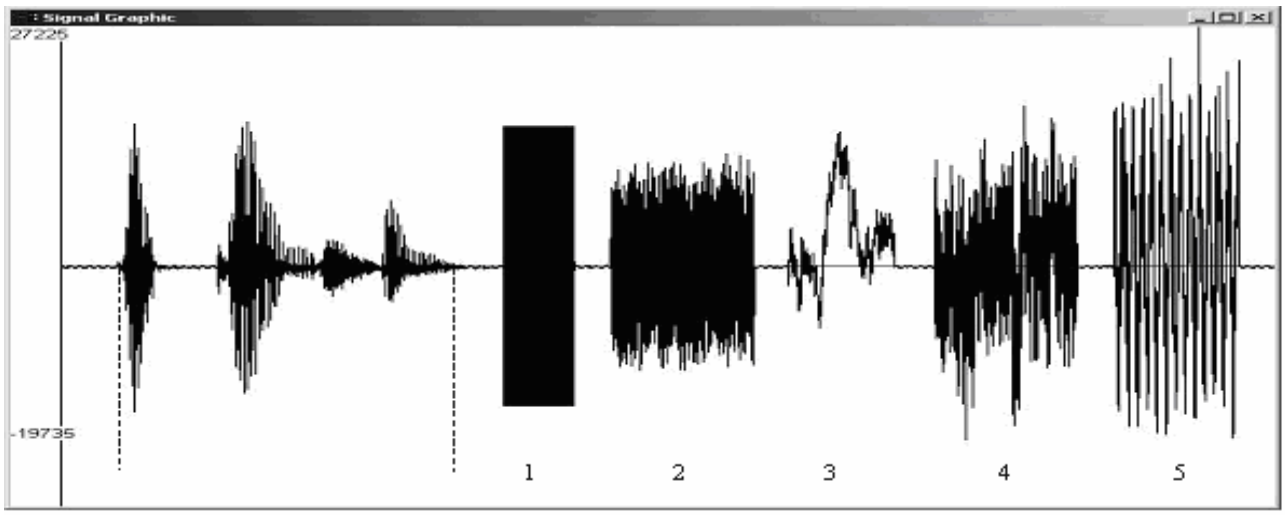

Рис. 3.4. Тестовый сигнал 1 (речь и искусственные шумы)

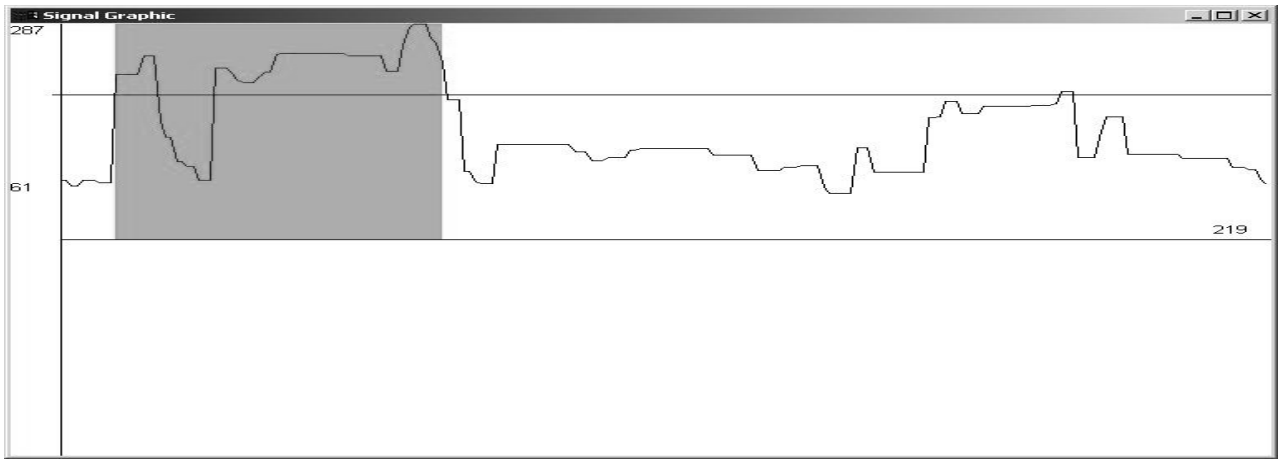

Рис. 3.5. Результат определения границ речи для тестового сигнала 1

На рисунках 3.6 и 3.7 представлен второй пример работы по выделению произношения того же слова из сигнала, который получается путем аддитивного смешивания исходного звукового сигнала с псевдослучайным белым шумом большой амплитуды. В данном эксперименте отношение сигнал/шум (SNR) равняется приблизительно 3 ДБ. Рисунок 3.7 показывает результат вычисления функции энтропии спектра сигнала и обнаруженный речевой фрагмент в сигнале.

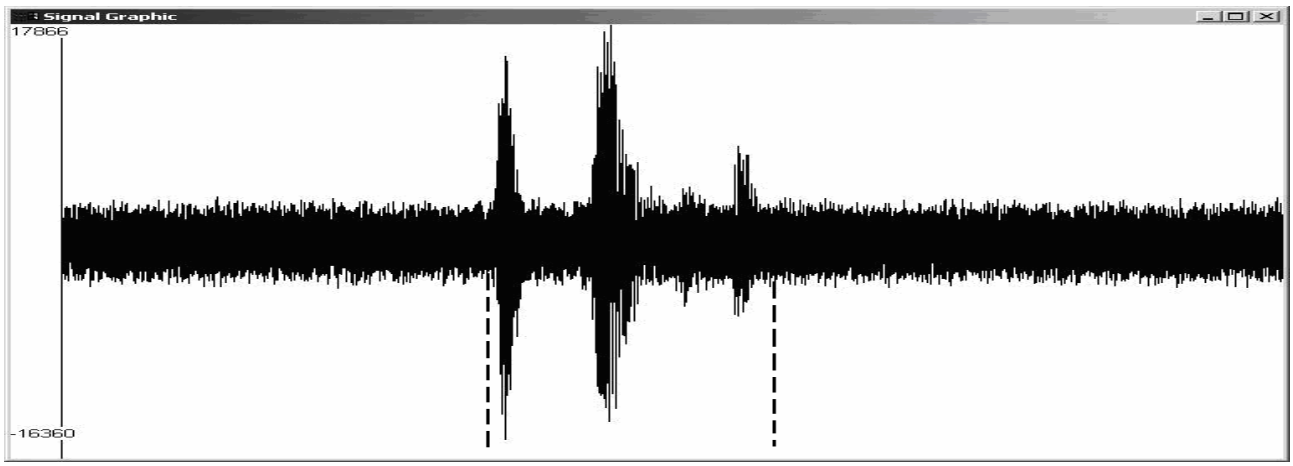

Рис. 3.6. Тестовый сигнал 2 (речь с аддитивным белым шумом)

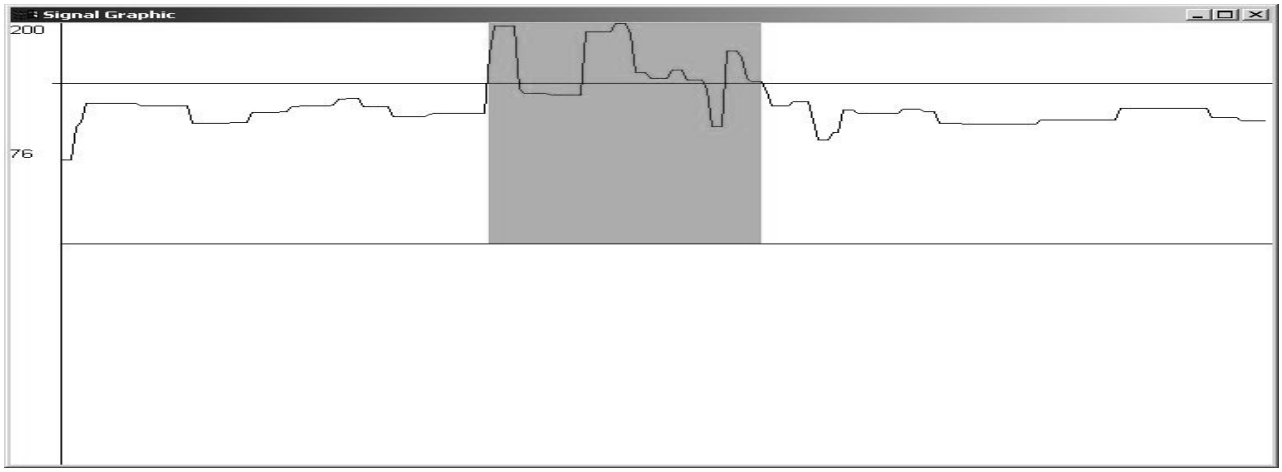

Рис. 3.7. Результат определения границ речи для тестового сигнала 2

При оценке работоспособности алгоритма может служить величина ошибки определения границ речи [23], которая складывается из двух показателей:

1) Вероятность ложной идентификации $P_{ложн.идент}$, т.е. определения факта наличия речевого фрагмента в момент времени, когда речевого сообщения на самом деле не было.
2) Вероятность пропуска речевого фрагмента $P_{проп.реч}$, т.е. отсутствия сигнала о наличии речевого сообщения в момент времени, когда речь в сигнале присутствует.

Таким образом, ошибка определения границ речи равняется:

$$P_{ош.гр.} = P_{ложн.идент.} + P_{проп.реч.} \quad (3.7)$$

В ходе экспериментов получены следующие результаты (Таблица 3.1) по автоматическому выделению изолированно произносимых слов в сигнале при наличии в нем искусственных шумов различного уровня.

Таблица 3.1. Процент ошибок определения границ речи

Вид шума		Вероятность ложной идентификации, %	Вероятность пропуска речи, %	Ошибка выделения речи, %
1	Узкополосный шум	0	2	2
2	Белый шум	1	2	3
3	Коричневый шум	3	2	5
4	Розовый шум	15	3	18
5	Усиленный акустический фон	2	2	4

Таблица 3.1 показывает высокую эффективность алгоритма при воздействии всех вышеперечисленных видов нестационарных случайных шумов (наихудший случай это присутствие розового шума, который по своей форме напоминает реальный речевой сигнал). Ошибка пропуска речевого участка, остается практически постоянной для всех экспериментов. Это объясняется периодическим усечением смычки перед взрывными согласными звуками ([т], [д], [п]) в начале слов, которая воспринимается как тишина.

Таким образом, экспериментальная проверка показала, что речевые фрагменты успешно выделяются из звуковых сигналов, содержащих сильные фоновые шумы. Кроме того, разработанный метод имеет достаточно высокую производительность, что позволяет его эффективно использовать в системах автоматического распознавания речи реального времени. Ограничения метода связаны с тем, что слабая фоновая речи или речеподобные сигналы могут быть приняты за полезную речь. Предложенный метод может быть полезен в

реальных системах, где возможны динамические изменения уровня громкости звукового сигнала в канале.

3.2. Выбор метода признакового описания речи

На этапе цифровой обработки сигнала непрерывный электрический сигнал проходит оцифровку и преобразуется в набор параметров, с которыми можно оперировать в дальнейшем. Основой задачей на этом этапе является получение компактных, но в то же время достаточно полно описывающих речевой сигнал характеристик. Наиболее эффективными системами признаков для задачи распознавания речи считают кепстральные признаки и коэффициенты линейного предсказания (КЛП). В группе речевой информатики СПИИРАН и лично автором также предложен метод параметризации речи, основанный на анализе формы спектра речевого сигнала [19]. Метод является развитием предыдущих идей [21] и отличается тем, что использует троичные признаки описания речи вместо двоичных признаков, что позволяет более точно описывать сигнал. В следующих разделах представлено описание разработанной системы признаков, а также результаты сравнения с существующими подходами к параметрическому представлению речевого сигнала.

3.2.1. Спектрально-разностные признаки речевого сигнала

Пусть имеется дискретный спектр сигнала $S = s_1, s_2, ..., s_n$. Чтобы избавиться от линейных деформаций спектра (изменения масштаба сигнала) достаточно использовать относительные коэффициенты $K_1, K_2, ..., K_{n-1}$:

$$s_1 = K_1 s_2, s_2 = K_2 s_3, ..., s_{n-1} = K_{n-1} s_n \quad (3.8)$$

Однако такое описание не решает проблему нелинейной деформации формы спектра. Для учета нелинейных деформаций введем для каждого участка аппроксимации спектра зону допустимых деформаций $\pm\alpha_i$ (рисунок 3.8 иллюстрирует эту зону).

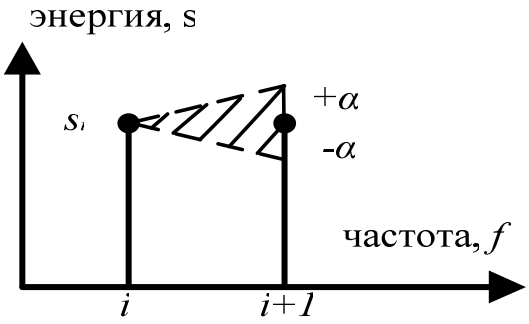

Рис. 3.8. Зона допустимых деформаций наклона спектральной функции речевого сигнала

Теперь перейдем от системы уравнений вида: $s_i = K_i s_{i+1} \pm \alpha_i$ к системе неравенств:

$$\begin{cases} s_i \geq K_i s_{i+1} - \alpha_i & (1) \\ s_i \leq K_i s_{i+1} + \alpha_i & (2) \end{cases} \qquad (3.9)$$

Система неравенств определяет один признак. Если выполняются оба этих условия, то формируется значение признака «приближенно равно». Если выполняется только условие (1) - значение признака «больше», если только условие (2) – значение признака «меньше». Одновременное невыполнение обоих неравенств исключено. Таким образом, каждый признак кодируется троичным кодом, например {0,1,2}.

На рисунке 3.9 представлен процесс преобразования речевого сигнала в вектора признаков.

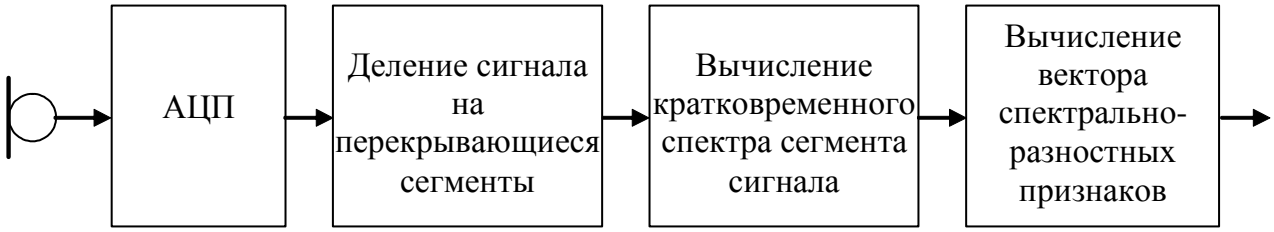

Рис. 3.9. Процесс вычисления вектора спектрально-разностных признаков речевого сигнала

Сначала речевой сигнал оцифровывается с частотой дискретизации 16 КГц и делится на короткие сегменты по 11 мс. Далее, используя алгоритм быстрого преобразования Фурье (БПФ) вычисляется кратковременный спектр сигнала. После этого формируется набор спектральных коэффициентов, характеризующих мощность сигнала на выходах гребенки фильтров. Для этого используется набор из 8 перекрывающихся треугольных фильтров (рисунок 3.10), значения частот которых подобраны экспериментальным путем (таблица 3.2), основываясь на исследованиях формантного состава речи.

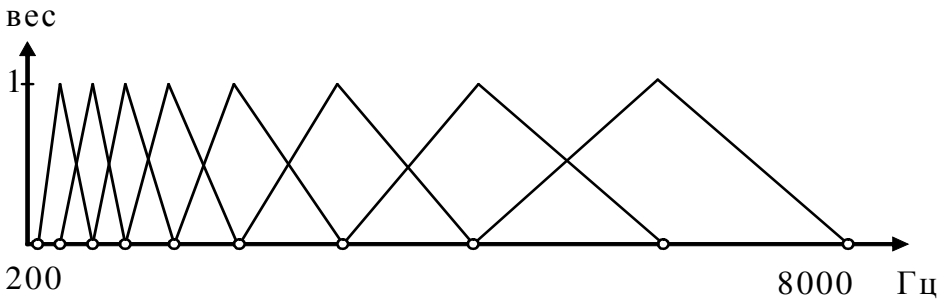

Рис. 3.10. Набор треугольных фильтров, применяемый к спектру сегмента речевого сигнала

Таблица 3.2. Параметры набора треугольных фильтров

Номер фильтра	Минимальное значение частоты, Гц	Максимальное значение частоты, Гц	Значение частоты вершины треугольника, Гц
1	200	500	300
2	300	750	500
3	500	1050	750
4	750	1500	1050
5	1050	2300	1500
6	1500	3400	2300
7	2300	6000	3400
8	3400	8000	6000

На последнем этапе обработки происходит знаковое сравнение различных пар спектральных коэффициентов. При этом, каждый признак кодируется двумя битами: «00» – возрастание спектральной функции на данном участке; «10» – спад спектральной функции на данном участке; «01» – отсутствие или малый наклон (попадающий в зону допустимых деформаций) спектральной функции на данном участке; код «11» не используется.

В проведенном исследовании система признаков получена полным перебором всех возможных пар спектральных коэффициентов с учетом относительных коэффициентов *K*: 1, 0.5, 2, 0.25, 4. Так при *n* спектральных полосах можно получить вектор, содержащий *C* троичных признаков для каждого соотношения:

$$C = \frac{n \cdot (n-1)}{2} \tag{3.10}$$

Таким образом, применяя 8 указанных выше спектральных полос и 5 значений коэффициентов, получаем вектор признаков, содержащий 140 троичных компонент.

3.2.2. Оценка систем параметрического представления речи

Были проведены эксперименты для оценки систем параметрического представления речи на задаче распознавания изолированных слов [19]. В ходе экспериментов разработанная система признаков сравнивалась с двумя другими системами спектральной природы: кепстральными (по mel-частотной шкале) и коэффициентами линейного предсказания. Кепстральные признаки использовались с модификациями: + признак энергии спектра сигнала, а также + производные, взятые от кепстральных коэффициентов.

Результаты сравнения представлены на рисунке 3.11. Из рисунка видно, что наилучшую точность распознавания обеспечивают mel-частотные кепстральные коэффициенты, состоящие из 13 коэффициентов с их первой и второй производными, хотя добавление второй производной практически не улучшило качество распознавания.

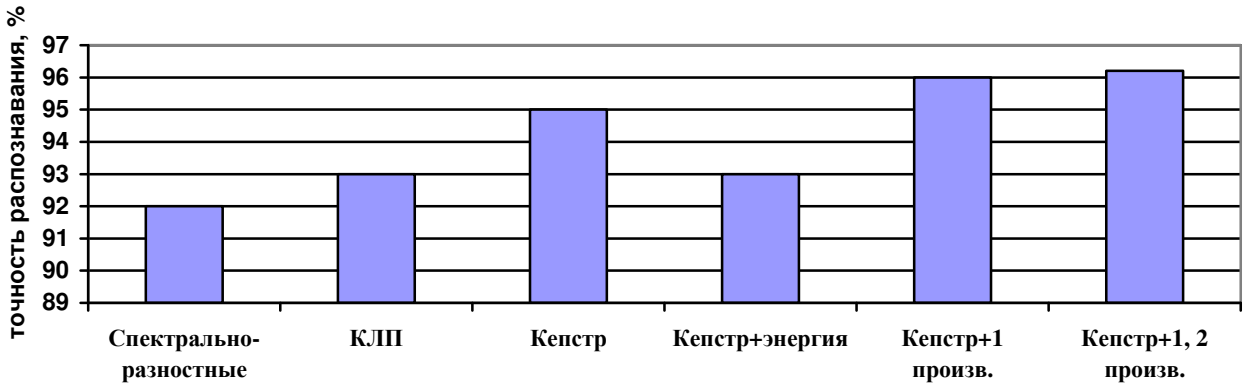

Рис. 3.11. Сравнение систем параметрического представления речевого сигнала по точности распознавания

Кроме проверки точности распознавания в работе была оценена робастность систем признаков к вариациям уровня громкости сигнала, которые являются наиболее частой причиной ошибок. В данном эксперименте мера робастности определяется, как допустимое значение неблагоприятных факторов, при котором система сохраняет свои характеристики [28], т.е. обеспечивает точность распознавания не ниже заданной.

На рисунке 3.12 приведены результаты проведенных экспериментов по оценке робастности систем признаков и сравнение их результатов с аналогичными экспериментами, выполненными для кепстральных признаков.

В проведенном эксперименте вначале была оценена точность распознавания при подаче на вход распознавателя исходного варианта тестовых произношений. Уровень громкости первоначального произнесения был принят за 100%. Затем были произведены эксперименты, в которых эталонный набор оставался неизменным, а уровень громкости воспроизведения исходного тестового набора последовательно понижался. Были произведены тесты при 2-х, 4-х и 8-ми кратном ослаблении уровня громкости сигнала в 2, 4 и 8 раз от первоначального уровня. Эксперимент показал, что спектрально-разностные признаки обеспечивают высокую точность распознавания, несмотря на многократное ослабление сигнала.

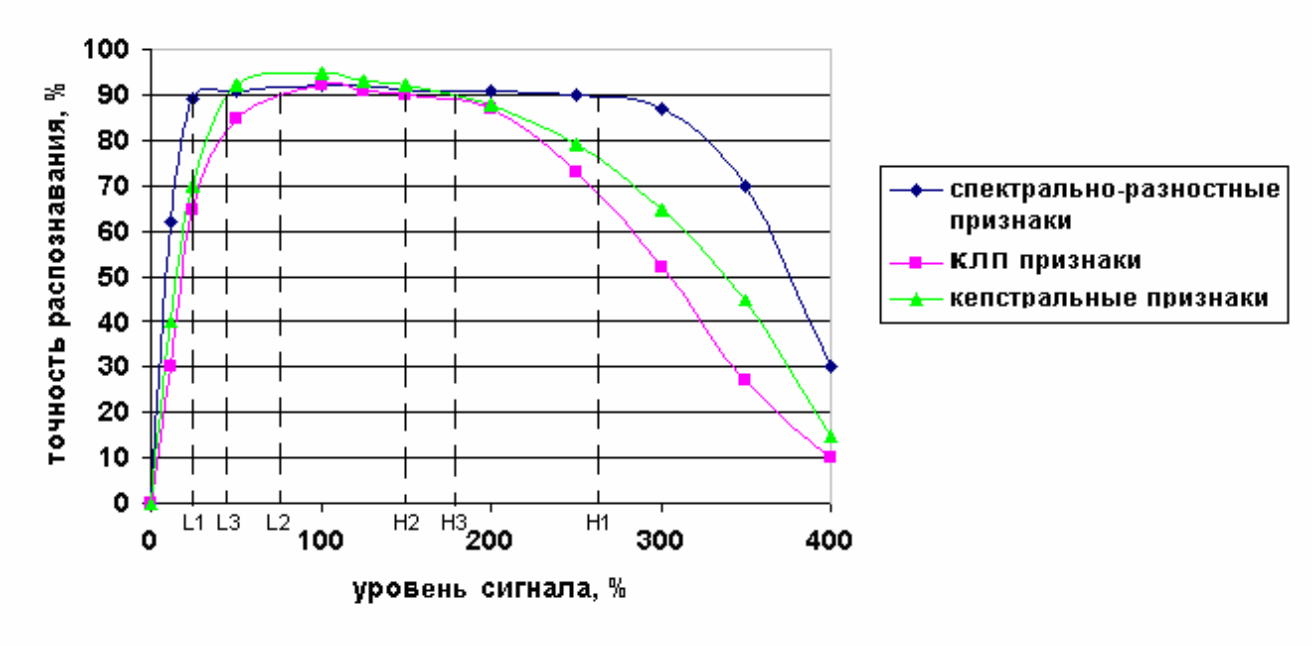

Рис. 3.12. Сравнение систем параметризации речи по критерию робастности к изменениям уровня громкости сигнала

Далее был проведен ряд аналогичных экспериментов, но с постепенным увеличением громкости воспроизведения тестового набора в 1.25, 1.5, 2, 3 и 4 раза. Необходимо отметить, что при увеличении уровня громкости сигнала наблюдается процесс отсечки сигнала по максимально допустимому уровню при 16-битном представлении сигнала. Несмотря на это, система оказывается устойчивой и к увеличению громкости сигнала. Причем, для спектрально-разностных признаков значительное снижение качества распознавания наблюдалось лишь при усилении сигнала более чем в 3 раза.

Зоной робастного распознавания речи принят уровень максимального усиления (ослабления) исходного сигнала, при котором достигается точность распознавания слов свыше 90%. Анализируя графики на рисунке 3.12 можно получить следующие границы робастного распознавания речи: для спектрально-разностных признаков нижняя граница уровня сигнала L_1=20%, верхняя граница уровня сигнала при робастном распознавании речи H_1=270%, для коэффициентов линейного предсказания те же параметры L_2=40%, H_2=180%

и для кепстральных признаков L_3=70%, H_3=160%. Таким образом, зона робастного распознавания речи будет равняться:

$$R = H - L \tag{3.11}$$

Зона робастного распознавания для спектрально-разностных признаков R_1=250, для автокорреляционных признаков R_2=90, и для кепстральных признаков R_3=140. Сравнивая данные значения можно сделать вывод, что робастность спектрально-разностные признаков к изменениям уровня громкости сигнала намного выше, чем для других методов параметризации.

Таким образом, в ходе экспериментов было выяснено, что кепстральные признаки обеспечивают несколько более высокую точность распознавания речи, однако предложенные спектрально-разностные признаки оказываются более робастными, что особенно важно в реальных условиях эксплуатации, когда может значительно меняться уровень входного сигнала в процессе речевого взаимодействия.

3.3. Метод распознавания русской слитной речи с включением морфемной обработки языка и речи

Процесс распознавания представляет собой метод оценки вероятности принадлежности речевого сигнала той или иной последовательности состояний модели. В ходе распознавания неизвестный входной сигнал представляется в виде параметрического описания O и оценивается моделью в соответствии с формулой:

$$v = \arg \max_{Ph} [P(O \mid \lambda_{Ph})] \tag{3.12}$$

где $1 \le Ph \le V$ является моделью фразы (гипотезой фразы), состоящей из СММ слов и фонем. При этом при большом словаре распознавания общее количество всех фраз оказывается несоразмерно большим. Поэтому для более адекватного моделирования русского языка и речи в модель распознавания речи введен дополнительный уровень обработки - морфемный. Рисунок 3.13 показывает уровни моделирования языка и речи в предложенной модели (возможно

сравнение с рисунком 1.7 базовой модели распознавания речи). За счет разделения слов языка предметной области на морфемы словарь распознаваемых языковых единиц значительно сокращается, но как показано во второй главе, он покрывает большее пространство слов языка. Модификация метода распознавания слитной речи за счет добавления морфемного уровня обработки не столь значительна как модификация методов обучения. Тем не менее, дополнительный уровень обработки вносит свои особенности в поздние уровни автоматического распознавания речи (см. рисунок 3.1).

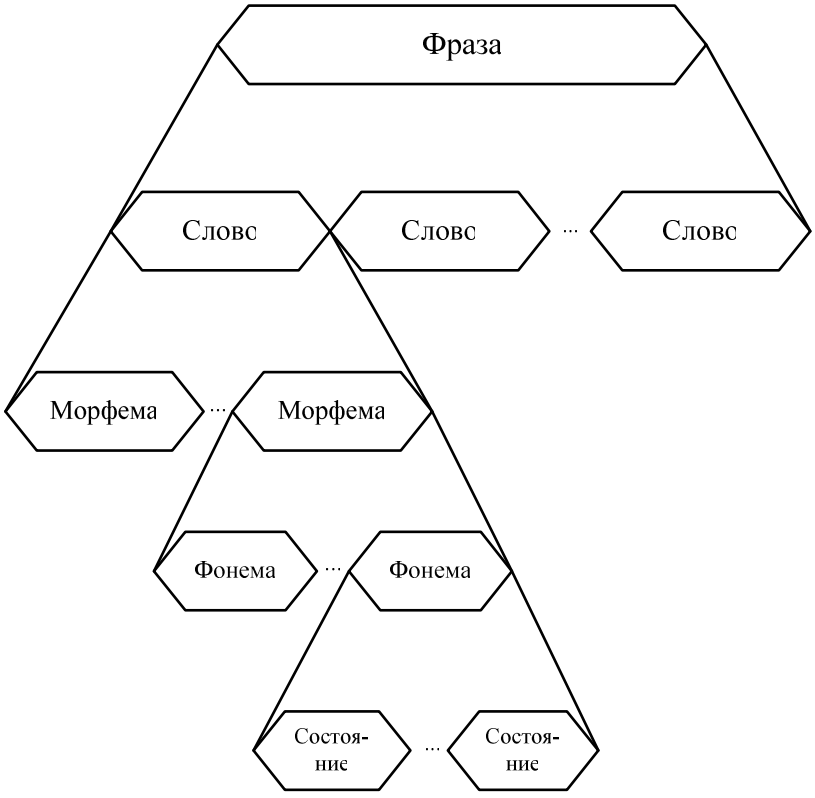

Рис. 3.13. Уровни моделирования языка и речи в модели

Метод передачи маркеров используется на уровне распознавания фонем и морфем. Причем, уровень распознавания фонем не претерпел изменений и используется метод передачи маркеров, описанный в разделе 1.3.4 и позволяющий определить оптимальную последовательность состояний, наилучшим образом соответствующих последовательности векторов наблюдений *O*. Но на следующем уровне распознавания модели фонем

объединяются не в модели слов, а в модели морфем с учетом межморфемных переходов, задаваемых морфемной моделью языка, и таким образом, объединенная СММ имеет вид, показанный на рисунке 3.14.

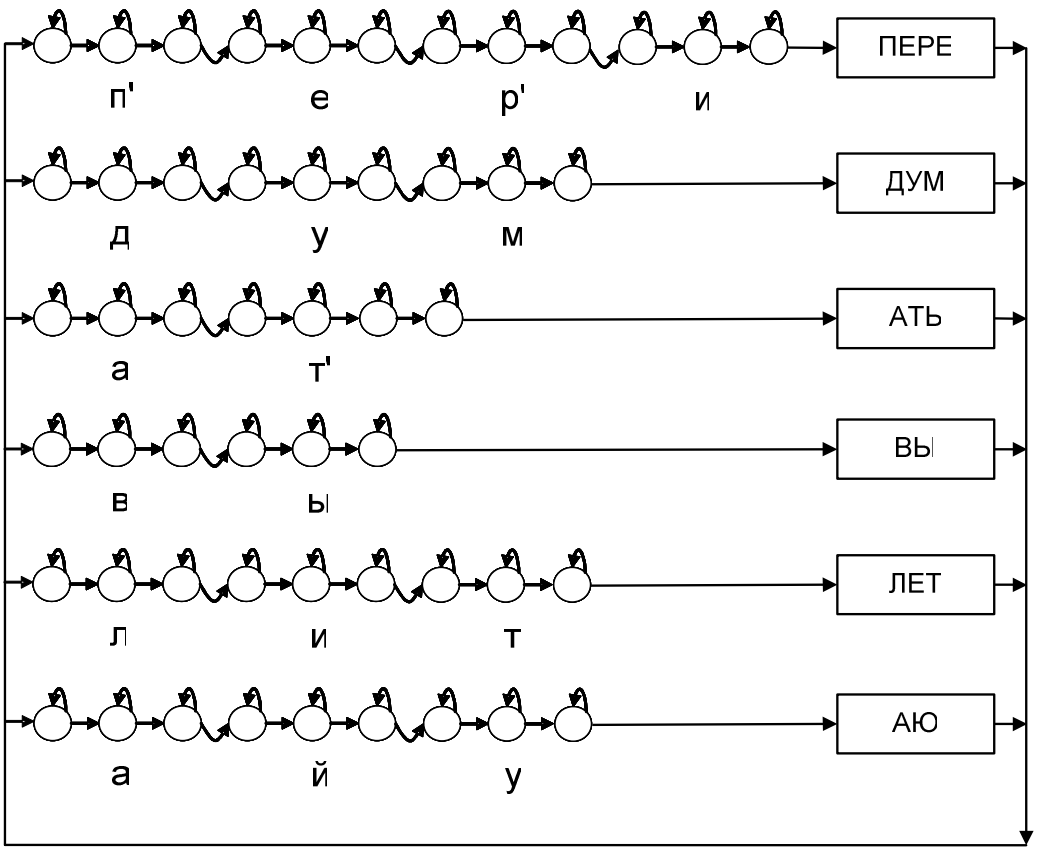

Рис. 3.14. Пример распознающей модели на основе морфемного представления языка прикладной области

Рисунок 3.14 поясняет простой пример, демонстрирующий преимущество морфемного моделирования языка при распознавании русской речи. Допустим, в ходе обработки некоторого обучающего текста встречены два слова: «передумать» и «вылетаю». Тогда, используя метод морфемного анализа, каждое из слов декомпозируется на 3 морфемы: «пере», «дум», «ать» и «вы», «лет», «аю» и словарь распознавания будет состоять из 6 данных языковых единиц. Однако в процессе распознавания их комбинации также могут составлять большое количество других корректных словоформ, например: думать, думаю, передумаю, выдумать, выдумаю, перелетаю, перелетать, летаю,

летать, вылет, вылетать, дум, вы и т.д. Таким образом, даже в таком небольшом примере, всего 6 морфем позволяют распознать в 3 раза больше слов, что доказывает эффективность такой модели.

После уровней распознавания фонем и подбора наиболее вероятных цепочек морфем получившаяся лучшая гипотеза (или набор из нескольких лучших гипотез) используется далее для формирования цепочек слов. Этот процесс может быть представлен как некоторая процедура (синтеза) цепочки распознанных морфем $M = m_1, m_2, ... m_D$ во фразу *W,* состоящую из *K* слов:

$$Y : M \rightarrow W = w_1, w_2, ..., w_K \qquad (3.13)$$

В ходе исследований были опробованы несколько морфемных моделей распознавания речи. Первая из предложенных моделей использовала метод декомпозиции слов на морфемы 5 типов: приставка, корень, интерфикс, суффикс и окончание. При этом количество используемых морфемных единиц было достаточно мало и их средняя длина была невелика (менее 3 символов). При этом на этапе распознавания для синтеза слов из различных типов морфем была предложена схема согласования типов морфем, представленная на рисунке 3.15. В данной модели заданы начальное и конечное состояния внутри некоторого слова, а в остальных узлах присутствуют возможные типы морфем. Дугами обозначены возможные переходы, а числа означают максимальное количество допустимых переходов из состояния в состояние.

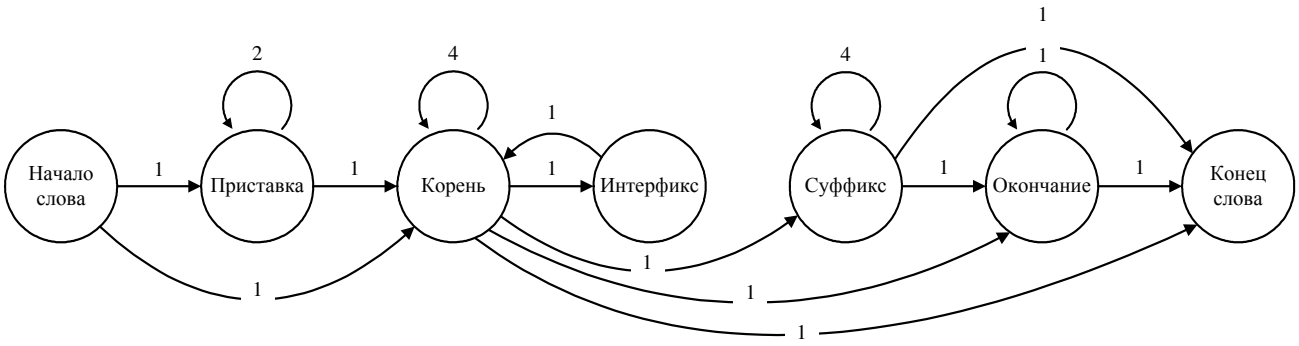

Рис. 3.15. Синтез слова из цепочки морфем в первом варианте морфемной модели распознавания

Однако данная модель синтеза слов из морфем на практике оказалась недостаточно эффективной, так как неизбежные ошибки распознавания на уровне фонем и морфем приводят к дополнительным ошибкам на уровне синтеза слов из-за неоднозначности в определении типа морфем. Короткие морфемы, особенно в интерфиксы, суффиксы и окончания, часто путаются между собой, поэтому не всегда можно однозначно определить тип морфемы, что приводит иногда к неправильному синтезу цепочки слов. Кроме того из каждой поступившей гипотезы фразы, представленной в виде последовательности морфем, можно синтезировать несколько различных гипотез фраз, представленных последовательностями гипотез слов. Частичное устранение этой неоднозначности возможно за счет использования морфологического анализа полученных гипотез, однако это значительно замедляет процесс обработки, так как необходимо использовать существующие грамматические и морфологические словари русского языка.

Для того чтобы избежать ошибок на уровне слов был использован упрощенный метод морфемного анализа и декомпозиции слов на морфемы (и псевдоморфемы), представленный в разделе 2.4. При использовании такого метода значительно упрощается процедура композиции слов из потока морфем так как применяется всего 3 типа морфем и при этом каждая морфема из распознаваемого словаря имеет пометку ее типа (1 - приставка, 2 - корень, 3 - концовка).

На рисунке 3.16 приведен алгоритм синтеза цепочки слов (гипотезы фразы) из цепочки распознанных морфем *m* количеством *D*. Основной проблемой является нахождение границ слов в потоке морфем, т.е. нахождение морфем, которые являются последними морфемами слов, за которыми нужно поставить разделитель (пробел) во фразе. В цикле производится проверка каждой морфемы из гипотезы в соответствии со следующим условием:

$$\begin{cases} m_i = граничн.морфема, (m_i \in S_{оконч}) \vee (m_i \in S_{корн} \wedge m_{i+1} \in S_{корн}) \vee (m_{i+1} \in S_{прист}) \\ нет, иначе \end{cases}, \quad (3.14)$$

где $S_{оконч}$ – множество концовок, $S_{корн}$ – множество корневых морфем и $S_{прист}$ – множество префиксов языка предметной области. Причем данные множества на этапе обучения выбираются таким образом, что подчиняются условию:

$$S_{прист} \cap S_{корн} \cap S_{оконч} = \varnothing \qquad (3.15)$$

Когда найдена граничная (последняя) морфема в синтезируемом слове, после него ставится разделитель и продолжается обработка оставшейся части морфемного потока.

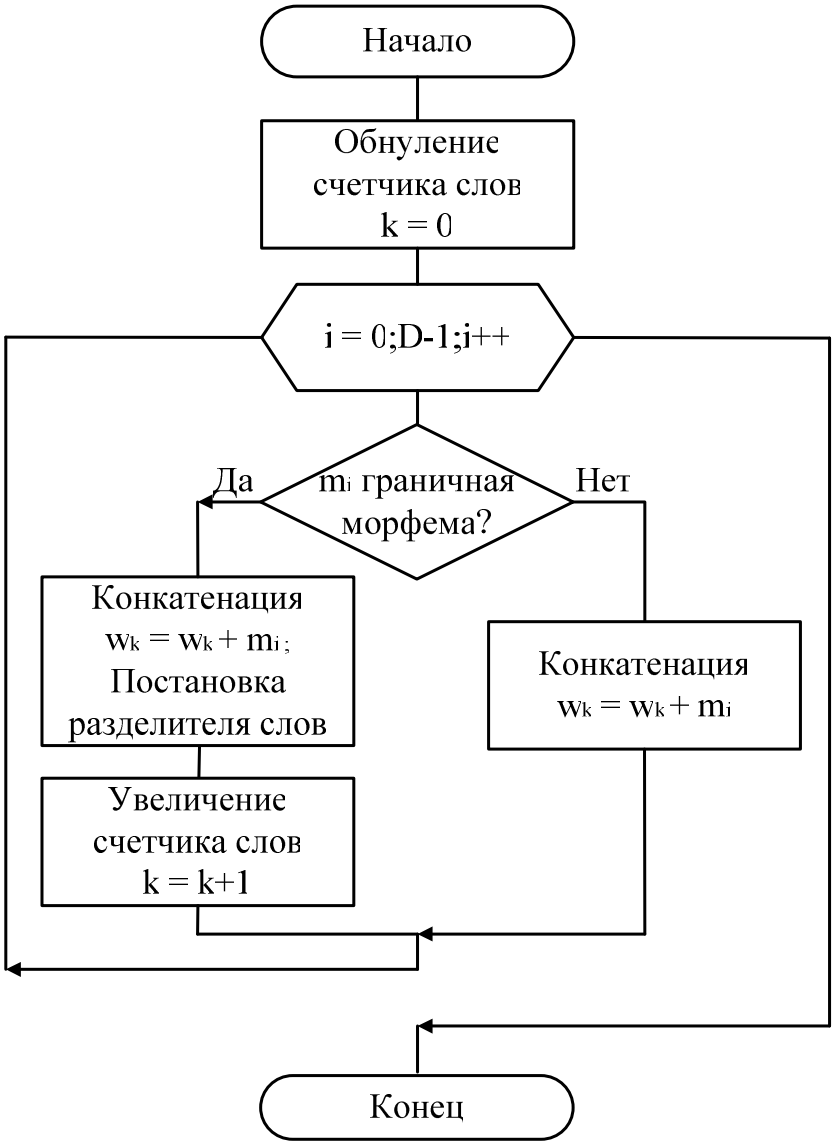

Рис. 3.16. Алгоритм композиции слова из потока морфем

Рисунок 3.17 поясняет процесс стыковки типов морфем для композиции слова. Такая модель композиции слов позволяет работать как с правильными

последовательностями морфем (образующими корректные слова), так и с допустимыми для языка последовательностями, имеющими ошибки на уровне морфем.

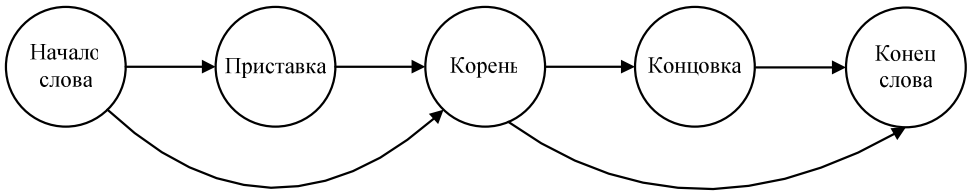

Рис. 3.17. Синтез слова из цепочки морфем во втором варианте морфемной модели распознавания

Дополнительным преимуществом предложенной модели синтеза слов перед первой моделью является ее скорость, так как в данном случае для определения границ слов достаточно информации, содержащейся в гипотезе фразы и никаких дополнительных словарей для постобработки текста гипотез фраз привлекать уже не требуется.

На выходе предложенной модели распознавателя слитной русской речи выдается гипотеза фразы, представляющая собой цепочку слов с разделителями между словами, которая затем может использоваться в системе понимания речи для определения смысла фразы.

Таким образом, представленная в третьей главе модель распознавания русской речи опирается на морфемную обработку, используя морфемную модель языка и морфемный словарь языка предметной области, создаваемые предварительно модулем обучения. Результаты экспериментов с моделью распознавания для прикладных задач и осуществленные внедрения модели приведены в четвертой главе.

Выводы по главе 3

1) Предложена архитектура модели автоматического распознавания русской слитной речи с морфемным уровнем обработки языка и речи, учитывающая механизмы словообразования и морфологические характеристики лексики русского языка и позволяющая существенно ускорить процесс распознавания за счет сокращения размера словаря и моделирования русского языка на уровне морфем.
2) Произведена модификация метода выделения полезной речи в звуковом сигнале за счет добавления логико-временного анализа функции спектральной энтропии сигнала. Проведенные эксперименты показывают, что данный метод позволяет выделять даже тихую речь в условии сильных акустических шумов.
3) Предложен метод троичного спектрально-разностного параметрического представления речевого сигнала, описывающий форму спектра сегментов сигнала, путем знакового сравнения значений энергии на выходах набора фильтров. В ходе экспериментов по сравнению систем признаков наилучшие результаты по критерию точности распознавания показали Mel-частотные кепстральные коэффициенты, а предложенные спектрально-разностные оказались более робастны к изменениям уровня сигнала.
4) Разработан метод композиции (синтеза) слов из цепочек распознанных морфем в процессе автоматического распознавания русской слитной речи, для чего определена функция нахождения граничных морфем слов в потоке морфем и формирования гипотезы распознавания фразы в виде цепочки слов с разделителями.

Глава 4. Программная реализация модели распознавания русской речи

Вторая и третья главы диссертации представляли исследования, направленные на создание теоретического фундамента модели дикторонезависимого распознавания русской слитной речи с большим словарем. Все предложенные автором модели и методы были реализованы программно и опробированы в нескольких прикладных задачах, где необходимо автоматическое распознавание речи. Были реализованы две модели, использующие функции автоматического распознавания речи: модель голосового доступа к справочному каталогу и модель бесконтактного управления компьютером, представленные в разделах 4.2 и 4.3 соответственно.

4.1. Архитектура программной реализации модели распознавания русской речи SIRIUS

В диссертационной работе была разработана программная реализация модели автоматического распознавания русской слитной речи. Данная реализация получила название SIRIUS (имеет расшифровку SPIIRAS Interface for Recognition and Integral Understanding of Speech) [18, 38]. Архитектура программной реализации SIRIUS представлена на рисунке 4.1. В данной реализации используется морфемный уровень представления и обработки русского языка и речи. Разделение словоформ языка предметной области на морфемы позволяет значительно сократить словарь распознаваемых языковых единиц, так как в процессе словообразования часто используются одни и те же морфемы. На основе правил словообразования русского языка и морфологического анализа созданы методы автоматической обработки текстов, а также базы данных морфем различных типов.

Все модули программной реализации разработаны для операционной системы семейства Microsoft Windows на языке программирования C++ [64] с

использованием методов объектно-ориентированного программирования в интегрированной среде разработки Microsoft Visual C++ 6.0.

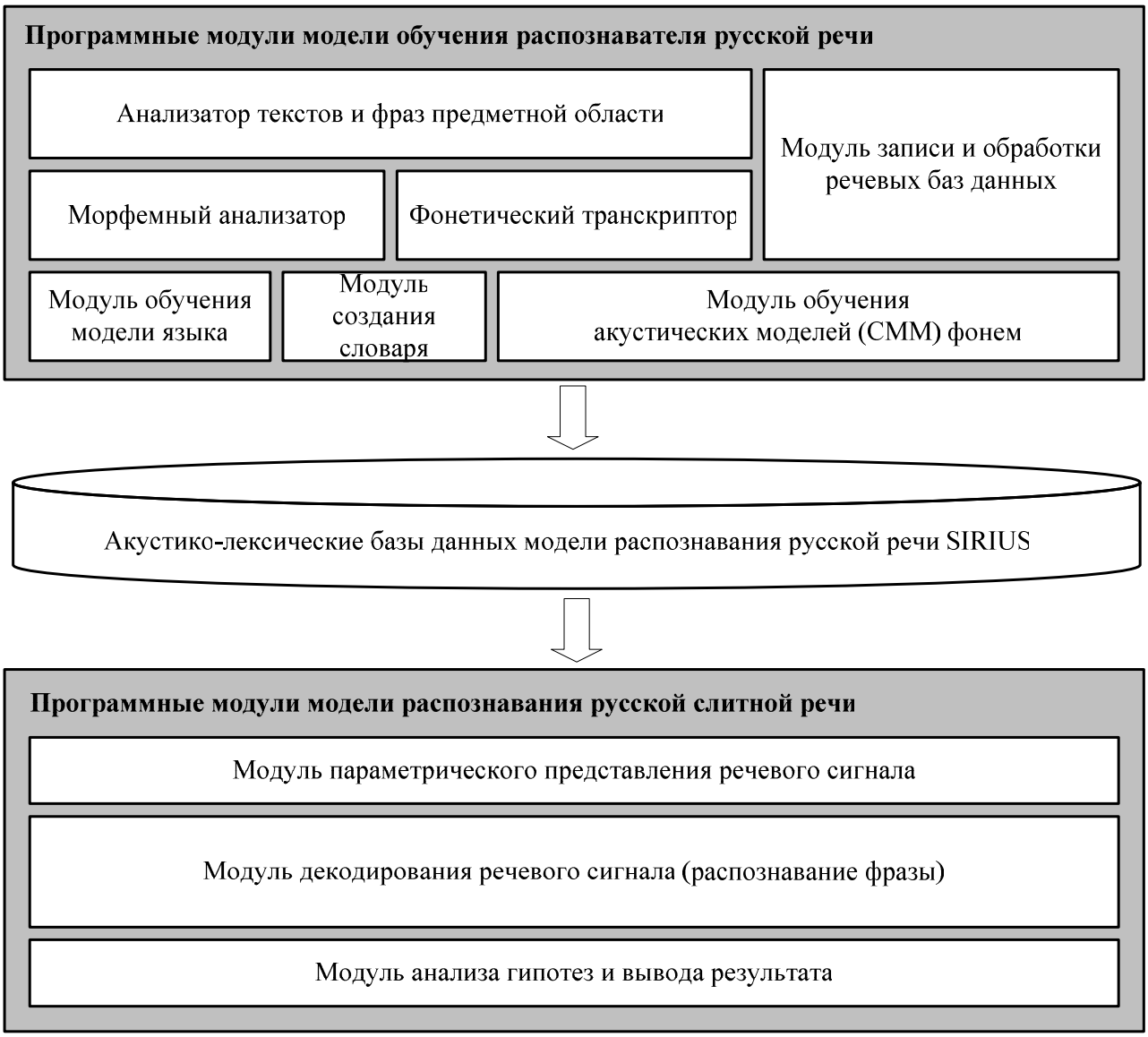

Рис. 4.1. Архитектура программной реализации модели распознавания русской слитной речи SIRIUS

Представленная архитектура реализует два этапа работы модели: обучение и распознавание речи. В результате работы модели обучения создаются акустико-лексические базы данных языка предметной области, которые затем используются модулем распознавания русской речи в ходе обработки звукового сигнала. Помимо этих основных моделей программная

реализация содержит модули для записи и обработки речевых баз данных, а также модуль анализа гипотез и результатов распознавания речи. Далее приводится описание основных функций и возможностей модулей (являются динамическими библиотеками DLL) программной реализации SIRIUS.

1) Программные модули модели обучения распознавателя русской речи:

- Модуль анализа текстов и фраз предметной области TextAnalysis.dll

Функция int TextAnalysis (char *pTextIn, char *pTextOut) служит для первичной области текстового корпуса, удаляя их него символы пунктуации, служебные символы и иностранные слова.

- Модуль морфемного анализа текста MorphAnalysis.dll

Функция int MorphAnalysis (char *pTextIn, char *pTextOut , char *pMorphologicalVocab, char *pMorphemeVocab) служит для морфемной декомпозиции слов исходного текста. Сохраняет разметку текста на фразы, добавляя разметку слов на морфемы. Использует для анализа слов текста морфологическую базу данных и морфемный словарь.

- Модуль фонетического транскрибирования PhoneticAnalysis.dll

Функция int PhoneticAnalysis (char *pTextInMorph, char *pTextOut) реализует набор правил фонетического преобразования текста. Представляет текст в виде фонетической транскрипции, используя для этого символы фонетического алфавита.

- Модуль записи и обработки речевых баз данных SpeechRecord.exe

Представляет собой исполняемую программу с графическим интерфейсом (диалоговое окно программы представлено на рисунке 4.2). Позволяет задать различные параметры звукового сигнала: частоту дискретизации, разрядность, максимальную длительность сигнала. Последовательно показывает фразы (в поле “Text of phrase”), которые необходимо произнести пользователю по нажатию кнопки “Record” в диалоговом окне. Остановка записи осуществляется либо по нажатию кнопки “Stop” либо по истечению установленного максимального времени записи. Каждая фраза автоматически сохраняется в новом файле.

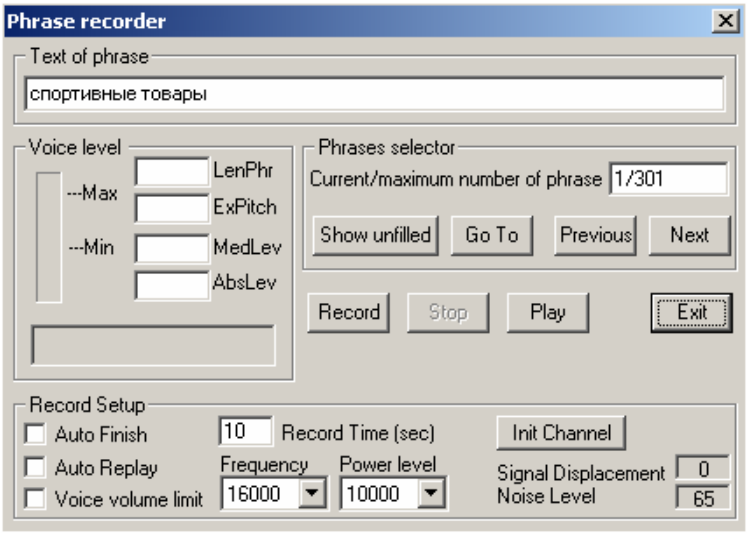

Рис. 4.2. Окно модуля записи и обработки речевых баз данных

- Модуль обучения статистической модели языка LangModel.dll

Функция int LangModelCreation (char *pTextInMorph, char *pFileLM) получает на входе обучающий текст с разделением слов на морфемы и создает на основе статистического анализа данного текста биграммную морфемную модель языка, сохраняя ее в текстовом файле pFileLM.

- Модуль создания словаря распознавания MorphVocab.dll

Функция int MorphVocabCreation (char *pTextInMorph, char *pTextInPhonetic, char *pFileVocab) позволяет создать словарь распознаваемых единиц, на вход функции подается текст с морфемной декомпозицией слов и соответствующий ему текст в фонетическом представлении, в результате создается файл содержащий в себе перечень уникальных морфем из первого текста с их фонетическими транскрипциями.

- Модуль обучения акустических моделей фонем HMMCreation.dll

Функция int HMMCreation (char *pPhoneticAlphabet, char *pRecordsDir, char *pTextInPhonetic, char *pFileHMMsOut) инициализует и обучает СММ для каждой фонемы из фонетического алфавита, используя для этого множество речевых файлов *.wav, находящихся в каталоге pRecordsDir и соответствующие им транскрипции произнесенных фраз. Набор обученных СММ сохраняется в текстовом файле.

2) Программные модули модели распознавания русской слитной речи:

- Модуль предобработки сигнала и параметрического представления речи ParamCalculation.dll

Функция int ParamComputation (char *pWaveFilesDir, char *pParamFilesDir, char *pParamConfig, bool VadEnabled) позволяет вычислить признаки речевого сигнала. Обрабатывает *.wav файлы каталога pWaveFilesDir, сохраняя файлы векторов параметров *.par в каталоге pParamFilesDir. Если первый параметр равен NULL, то производится захват звукового сигнала с микрофона. pParamConfig задает тип вычисляемых параметров: “MFCC” – кепстальные признаки, “LPC” – коэффициенты линейного предсказания, “SD” – спектрально-разностные признаки. Дополнительно может принимать значения: “+E” – используется энергия сегмента, “+D” – используется первая производная коэффициентов, “+A” – используется вторая производная. VadEnabled=TRUE позволяет использовать детектор границ речи.

- Модуль декодирования речевого сигнала SIRIUS.dll

Функция int SIRIUS_Init (char *pFileHMM, char *pFileLM, char *pFileVocab) производит инициализацию распознавателя русской речи SIRIUS, загружая в память программы CMM фонем, модель языка и распознаваемый словарь.

Функция int SIRIUS_Recognition (char *pParamFile, char *pRecognResult, char *pParamConfig, char *pRecognLevel) запускает процесс распознавания сигнала, считываемого из *.wav (или *.par) файла либо, поступающего от микрофона (если pWaveFile=NULL). Результат распознавания сохраняется в текстовой переменной pRecognResult. pParamConfig задает тип вычисляемых параметров сигнала, pRecognLevel позволяет задать форму отображения результата: “P” – результат распознавания на уровне фонем, “M” – морфем, “W” – слов.

- Модуль анализа гипотез и вывода результата RecognResult.dll

Функция int ResultAnalysis (char *pResultIn, char *pResultOut, bool IsEndings) осуществляет постобработку гипотезы распознавания речи. IsEndings=FALSE позволяет удалить концовки всех слов гипотезы распознавания.

Функция int StatisticsCalc(char *pPhrasesFile, char *pResultsFile, double Accuracy) сравнивает оригинальные тексты произнесенных фраз и тексты распознанных фраз и сохраняет в Accuracy статистику точности распознавания в процентах.

4.2. Модель голосового доступа к электронному справочному каталогу

В рамках текущего инновационного проекта ИНТАС № 05-1000007-426 «Introduction of the automatic Russian speech recognition system SIRIUS in telecommunications» проводится внедрение программной реализации модели распознавания русской речи SIRIUS в инфотелекоммуникационные сервисы, разрабатываемые совместно с телекоммуникационной компанией «NewVoice» [45]. В частности, разрабатывается модель голосового заказа авиабилетов и модель голосового доступа к электронному справочному каталогу «Желтые страницы Санкт-Петербурга» с целью создания единой автоматизированной справочной системы для голосового ввода запросов адресов и телефонов организаций как при помощи компьютера со стационарным микрофоном, так и при помощи мобильного телефона. Электронной каталог «Желтые страницы», доступный пользователям сети Интернет [55], является полным телефонным справочником организаций и фирм города. Каталог содержит перечень всех организаций с указанием их адресов, телефонов и видов деятельности. Структура каталога организована в виде дерева, представленного на рисунке 4.3. Каталог периодически пополняется, но на начало 2006 года общее количество тематических названий рубрик в каталоге (N) – более 2050 с длиной фраз от 1 до 5 слов (например, «рестораны французской кухни»); общее количество фирм в каталоге (Z) – более 60000.

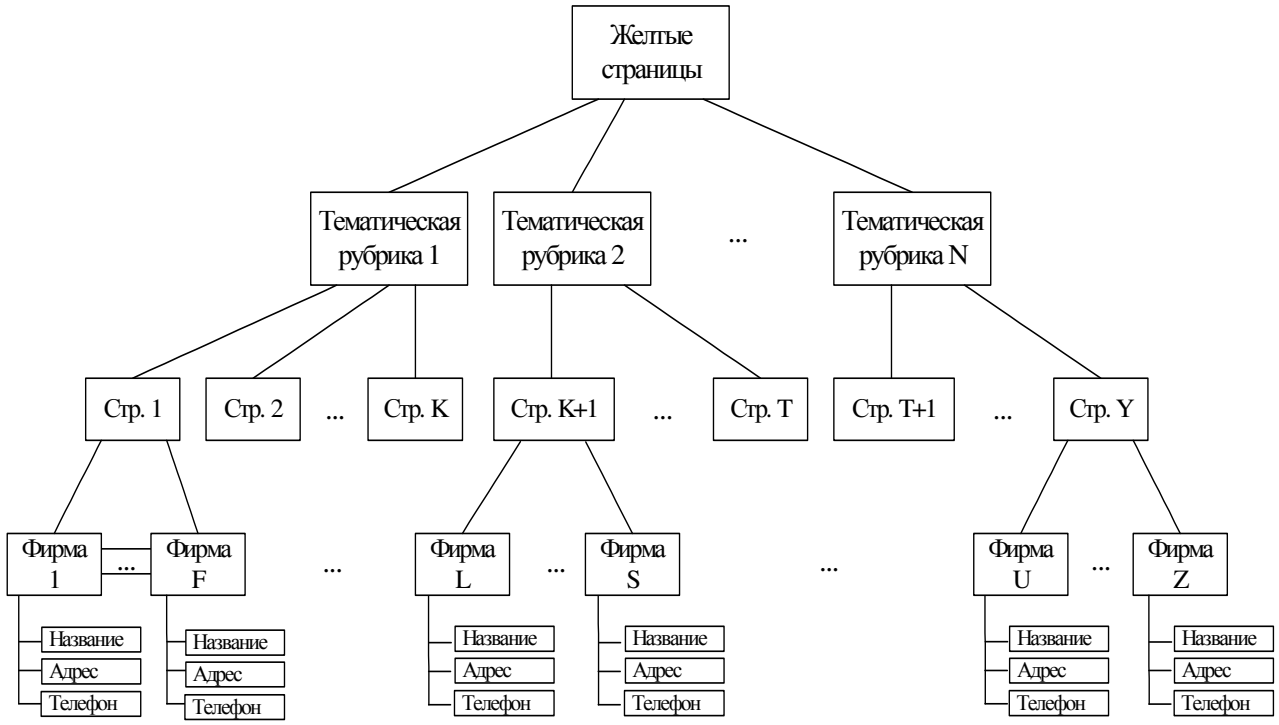

Рис. 4.3. Информационная структура электронного каталога «Желтые страницы»

Создание автоматизированной справочной системы с голосовым доступом к ресурсам электронного каталога сети Интернет посредством компьютера или телефона позволяет использовать для ответов на звонки абонентов не операторов-людей, а систему автоматизированного распознавания речи с генерацией речевых ответов. Такая система представляет собой программно-аппаратный комплекс, называемый голосовой портал [17]. Общая структура голосового портала представлена на рисунке 4.4. Он предоставляет пользователям возможность получать и управлять информацией, размещенной в глобальной сети Интернет, при помощи голоса. При этом пользователь имеет возможность использовать необходимый ему сервис как посредством телефона, так и голосового броузера (программного интерфейса для персонального компьютера).

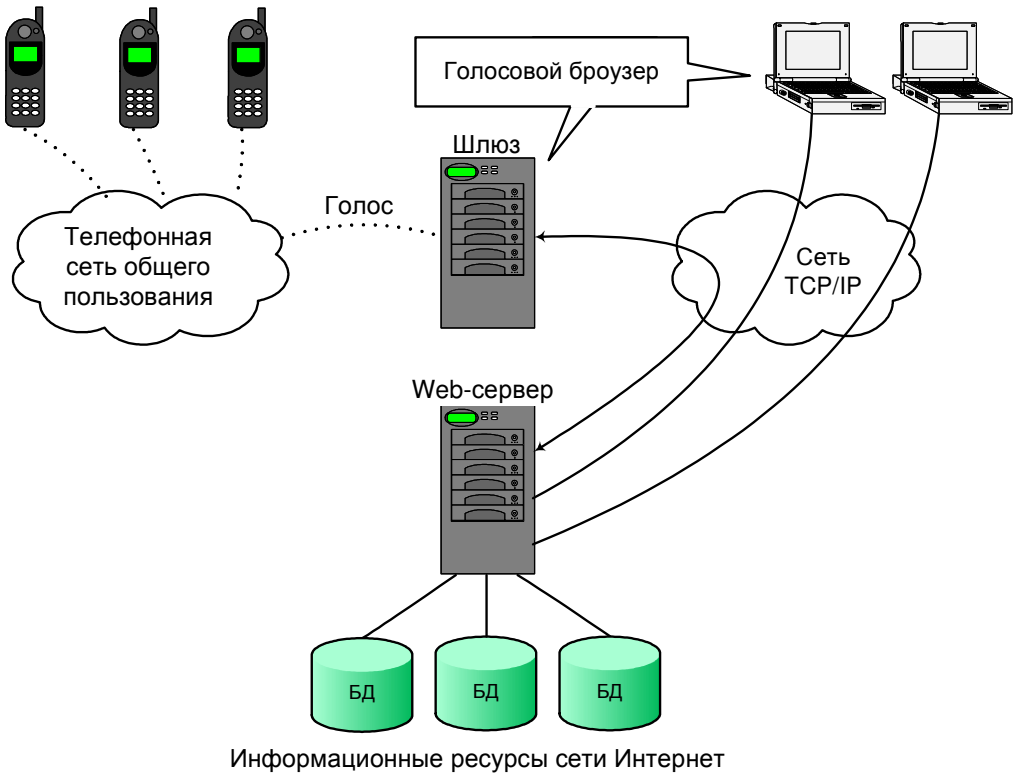

Рис. 4.4. Общая архитектура телекоммуникационного голосового портала

4.2.1. Описание модели голосового доступа к каталогу

В ходе работ реализована демонстрационная модель голосового ввода запроса и поиска названий рубрик в электронном справочном каталоге «Желтые страницы». Демонстрационная модель дает возможность пользователю получить перечень организаций, соответствующих искомой рубрике, при помощи микрофона, подключенного к персональному компьютеру. На рисунке 4.5 представлен пример работы модели голосового доступа к электронному каталогу. После голосового ввода пользователем фразы, модель распознает поступивший сигнал, выполняет запрос и отображает на экране информацию, соответствующую запросу.

Размер словаря в данной задаче составляет 1850 слов, а при разбиении слов на морфемы размер морфемного словаря сокращается до 1360 языковых единиц. Незначительное сокращение словаря связано со спецификой языка данной задачи. Во-первых, слова таких частей речи как глаголы, причастия и местоимения, имеющие особенно сложную структуру словообразования, почти

не употребляются в языке данной предметной области. Во-вторых, практически все существительные и прилагательные употребляются только в именительном падеже единственного или множественного числа.

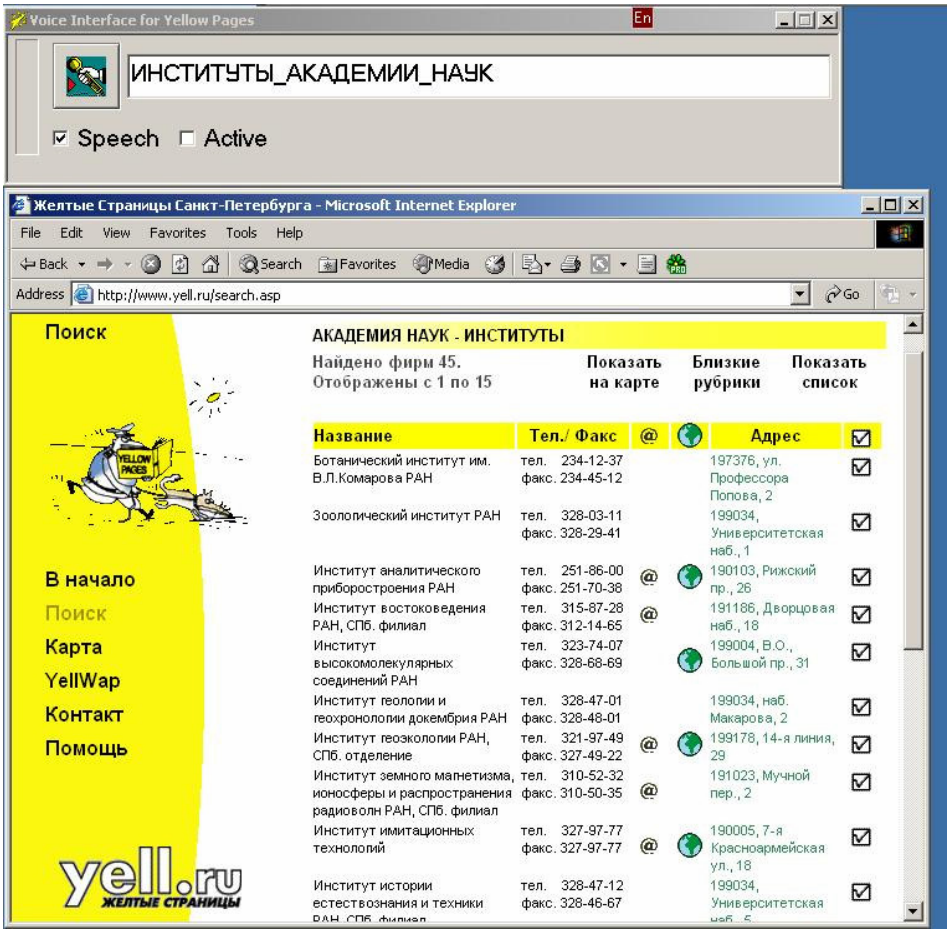

Рис. 4.5. Пример работы модели голосового доступа к электронному справочному каталогу

Для работы с моделью были записаны 25 мужских голосов дикторов (в возрасте от 20 до 30 лет), 500 фраз для каждого (названия тематических рубрик каталога). Звуковые файлы 20 дикторов использовались для задачи обучения модели и 5 оставшихся дикторов - для тестирования модели распознавания.

Созданный речевой корпус был использован для тестирования созданных в ходе работы моделей целословного распознавания русской речи, морфемного, слогового и фонемного распознавания речи. Результаты сравнения моделей по

точности распознавания приведены далее в разделе 4.2.2, а по скорости распознавания - в разделе 4.2.3.

4.2.2. Сравнение моделей распознавания русской речи по точности распознавания

Для выбора оптимальной стратегии распознавания русской речи были реализованы несколько экспериментальных моделей: целословная модель, морфемная, слоговая и фонемная. Оценка их моделей языка приводилась выше в разделе 2.5, где было показано преимущество моделирования русского языка посредством слоговой и морфемной моделей.

Результат сравнения четырех моделей по точности распознавания приведен на рисунке 4.6. Для целословной и морфемной моделей приведены точность распознавания фонем (столбец 1), точность распознавания слов (столбец 2) и точность распознавания фраз (столбец 3). Точность распознавания слов для слоговой и фонемной моделей посчитать затруднительно, так как не существует приемлемого решения для синтеза слов из распознанной цепочки слогов (или фонем). Проблема заключается в том, что не существует правил определения границ слов, имея цепочку слогов (или фонем). Тут следует также учитывать, что распознанная цепочка может быть искажена ошибками при распознавании. Границы слов для морфемной модели распознавания проставляются по принципу, который был описан выше в разделе 3.3.

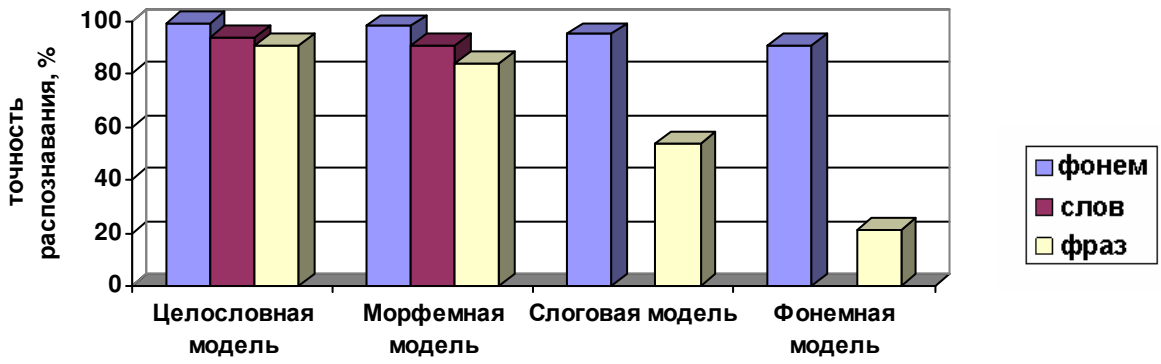

Рис. 4.6. Сравнение точности распознавания фонем, слов и фраз при использовании моделей распознавания

Вычисление точности распознавания речи A проводилось последовательно на каждом из уровней распознавания (фонем, морфем, слов и фраз) в соответствии со следующей формулой [115]:

$$A = 100\ \% \cdot (1 - \frac{S + D + I}{N}) , \qquad (4.1)$$

где S - количество неверно измененных языковых единиц; D - количество пропущенных языковых единиц; I - количество вставок лишних языковых единиц; N – общее количество языковых единиц в тестовых фразах.

Пример распознавания входной фразы «институты академии наук», используя каждую из моделей, приведен на рисунке 4.7. Курсивом показаны ошибки распознавания, сделанные системой. Как видно из рисунка из-за ошибок в распознавании могут возникать даже новые слова, искажающие смысл всего высказывания. Как уже было сказано выше, ни слоговая модель, ни фонемная не позволяют однозначно определять границы слов в потоке речи, поэтому их применение для распознавания слитной речи не эффективно.

Произнесенная фраза	институты_академии_наук
Целословная модель	институты_академи***я***_наук
Морфемная модель	институт2_ы3_академ2_и***я***3_наук2
Слоговая модель	ин_сти_ту***т***_ты_а_ка_де_ми_***ки***_на_ук
Фонемная модель	и_н_с’_т’_и_т_у!_***т***_т_ы_а_к_а_д’_е!_***н***_и_***к’***_и_н_а_у_к

Рис. 4.7. Примеры ошибок, сделанных моделями при распознавании фразы

Были проанализированы результаты и ошибки распознавания речи при применении морфемной и целословной моделей распознавания. Был произведен анализ классов ошибок распознавания слитно произнесенных фраз:

1) Замена одного слова другим

2) Пропадание слова

3) Вставка лишнего слова

4) Изменение словоформы внутри парадигмы слова

Было выяснено, что наиболее частой причиной ошибок является неправильное распознавание концовок словоформ (в рамках парадигмы одного и того же слова), которые произносятся обычно не так чётко как начала, а ошибки при распознавании слов приводят к тому, что происходит ошибка в распознавании всей фразы из-за несогласованности слов в предложении.

При этом в данной задаче искажение концовки словоформы не ведет к искажению смысла (в отличие от искажения корня или приставки), поэтому была модифицирована модель морфемного распознавания следующим образом: концовки слов не учитываются при распознавании, а принимаются во внимание только приставки и корни словоформ.

Для того чтобы можно было отбросить концовки слов из гипотезы распознавания словарь морфем системы распознавания речи и морфемная модель языка содержат пометки каждого типа морфемы (1 - пристака, 2 - корень, 3 - концовка), поэтому достаточно лишь в гипотезе распознавания отбросить морфемы типа «3».

Архитектура разработанной модели автоматического голосового доступа к электронному каталогу показана на рисунке 4.8. Для распознавания русской слитной речи используется программная реализация модели распознавания речи SIRIUS. Полученная в ходе распознавания лучшая гипотеза фразы проходит через модуль удаления концовок слов в тексте гипотезы, а затем модуль поиска названия рубрики, который сравнивает слова гипотезы со словами заранее подготовленного перечня названий рубрик электронного каталога. После чего формируется запрос к ASP скрипту поисковой страницы Интернет сайта электронного каталога и результат выполнения запроса выводится на экран.

Модифицированная модель, которая не учитывает концовки слов в гипотезе фразы, позволяет несколько повысить точность распознавания фраз по сравнению с базовой морфемной моделью распознавания. На рисунке 4.9

приводится сравнение базовой и модифицированной морфемной моделей по точности распознавания фраз (определения названия рубрики).

Рис. 4.8. Архитектура модели голосового доступа к электронному каталогу «Желтые страницы»

Точность распознавания фраз для модифицированной морфемной модели относительно базовой повысилась на 7% с 84.2% до 90.1% и практически достигла уровня целословной системы распознавания речи. При этом, как будет показано в следующем разделе, морфемная модель распознавания работает значительно быстрее целословной.

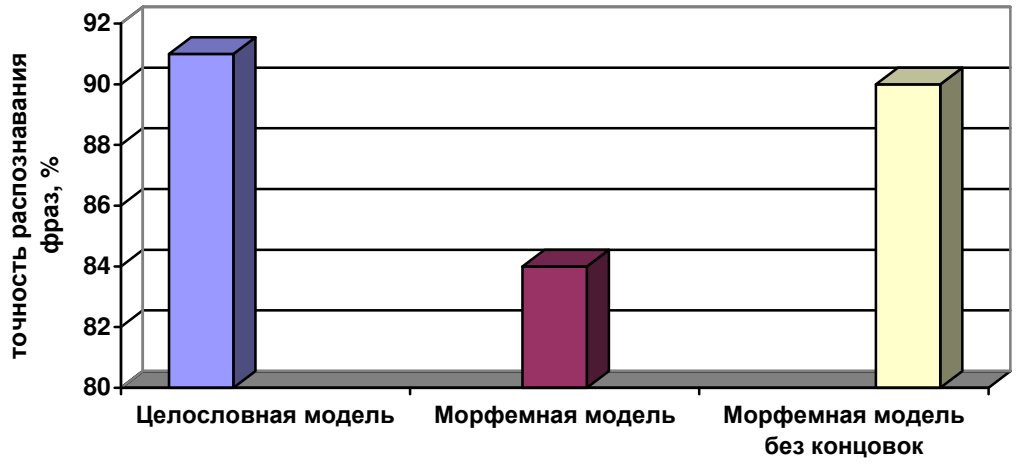

Рис. 4.9. Сравнение точности распознавания фраз для базовой и модифицированной морфемной моделей

4.2.3. Сравнение моделей распознавания русской речи по скорости обработки

При разработке диалоговых систем крайне важную роль играет время отклика системы на пользовательский ввод. Поэтому в ходе исследований были проведены эксперименты, направленные на сравнение скорости работы распознавателей, основанных на целословном, морфемном, слоговом и фонемном представлении языка прикладной задачи. Результаты экспериментов приведены на рисунке 4.10.

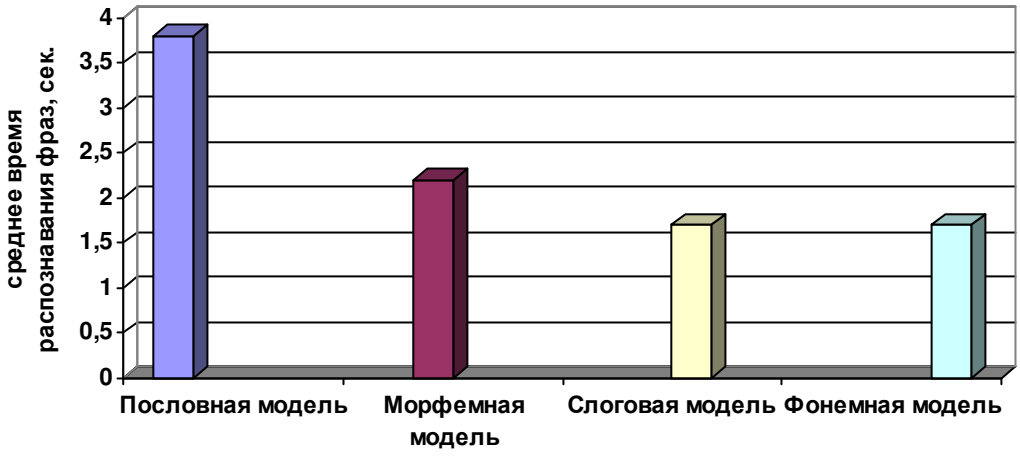

Рис. 4.10. Сравнение моделей распознавания по среднему времени обработки слитнопроизнесенных фраз

В случае ввода сигнала с микрофона, система, обнаружив наличие речи (методом анализа функции спектральной энтропии, описанным в разделе 3.1), сразу начинает процесс распознавания еще до окончания ввода фразы. Однако для сохранения одинаковых тестовых условий во всех экспериментах, в проведенных экспериментах тестовые фразы считывались из файла, а не вводились напрямую с микрофона. Средняя длина фразы по всем тестовым файлам составила 1.7 сек. (это значение принимается равным времени обработки в режиме реального времени (RT factor) [6] . Таким образом, время обработки речи при пословной модели распознавания равняется 2.24 RT (real-time), а при применении морфемной модели распознавания 1.28 RT, таким образом, улучшение составило 75%. При вводе же сигнала с микрофона задержка между окончанием ввода фразы и реакцией (ответом) модели распознавания составляет примерно 0.3 RT или 0.5 сек., что является приемлемым для диалоговых систем. При применении целословной модели распознавания время реакции автоматической системы будет свыше 2 секунд, что негативно воспринимается пользователями в человеко-машинном диалоге. Данные эксперименты проводились на персональном компьютере AMD Athlon 64 3400+, оснащенным 1 Гбайт оперативной памяти.

Слоговая и фонемная модели обеспечивают более высокую скорость распознавания, однако результаты по точности распознавания, приведенные в разделе 4.2.2, показывают неэффективность таких моделей по данному критерию. Предложенная же морфемная модель распознавания является оптимальной в смысле удачного компромисса между скоростью обработки речи и точностью распознавания слитной речи, т.е. модель в данной прикладной задаче модель обеспечила существенный рост производительности (свыше 75%) и минимальную потерю в точности распознавания (около 1%) по сравнению с целословной моделью. Эти результаты согласуются с результатами по автоматическому распознаванию речи, основанному на лексических единицах меньше слова, для ряда агглютинативных языков: турецкого [74], финского [85], корейского [96].

4.3. Модель бесконтактного управления компьютером

В ходе совместных научно-исследовательских работ группы речевой информатики СПИИРАН, проводимых по проекту EC-FP6 SIMILAR Network of Excellence [49], автором разработана одна из первых российских многомодальных моделей человеко-машинного взаимодействия, предназначенная для бесконтактного управления персональным компьютером (полностью без использования клавиатуры и мыши). Многомодальная модель получила название ICanDo (расшифровывается как Intellectual Computer AssistaNt for Disabled Operators) [91]. В 2006 году данная модель получила гран-при на Международном конкурсе многомодальных интерфейсов LoCo Mummy Contest 2006 в Брюсселе [47].

Предложенная модель необходима, в основном, для помощи людям, имеющим проблемы с двигательными функциями рук или же вообще без рук. Вместо клавиатуры и мыши для управления графическим интерфейсом пользователя в модели используются голосовые команды и движения головой. Используемая здесь программная реализация модели автоматического распознавания русской речи SIRIUS позволяет распознавать голосовые команды пользователя, которые выполняются в зависимости от контекста и текущего положения курсора.

4.3.1. Архитектура модели

В предложенной многомодальной модели используются две естественные входные модальности: речь и движения головы оператора. Так как обе модальности являются активными [99], то они непрерывно отслеживаются компьютером. Каждая из модальностей передает свою семантическую информацию: положение головы пользователя определяет положение курсора мыши в данный момент времени, а речевой сигнал передает информацию о действии, которое должно быть выполнено с некоторым объектом рабочего стола компьютера. На рисунке 4.11 представлена общая архитектура многомодальной модели бесконтактного управления компьютером.

Положение курсора мыши зависит только от положения нескольких отслеживаемых естественных точек на лице пользователя (кончик носа, левый глаз, правый глаз, точка между бровей, центр верхней губы [92]) и вычисляется непрерывно по мере обработки видеопотока [77]. В том случае, когда система распознавания речи зафиксировала и распознала некоторую голосовую команду, данную команду необходимо выполнить с учетом информации о положении курсора на экране монитора.

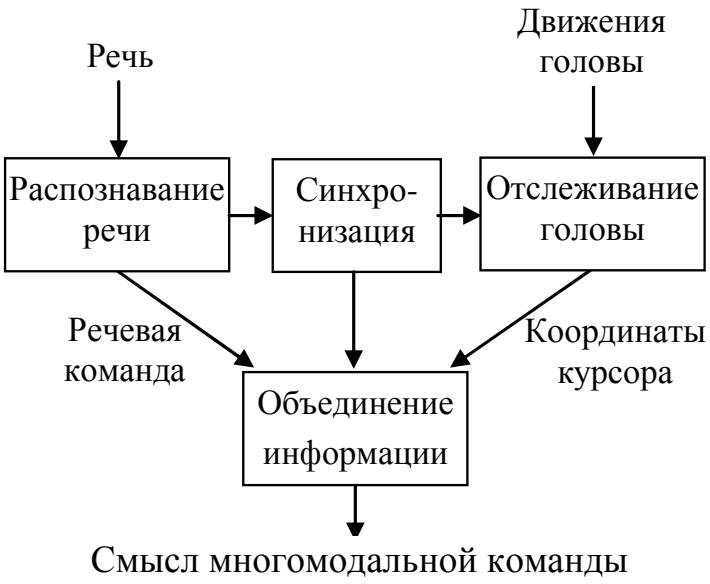

Рис. 4.11. Архитектура многомодальной модели бесконтактного управления компьютером

Рисунок 4.12 иллюстрирует процесс синхронизации многомодальных потоков в модели и объединения информации. На рисунке показан фрагмент выполнения сценария по бесконтактной работе с программой Internet Explorer для нахождения некоторой информации на Интернет сайте (последовательность голосовых команд «Левая», «Вниз» и «Левая»), копирования фрагмента этой страницы в буфер (команды «Нажать левую», «Отпустить левую» и «Копировать»), открытие редактора MS Word (команды «Пуск» и «Левая»), а также вставка информации из буфера в окно текстового редактора (команда «Вставить»). Черный кружок означает, что распознанная

команда (например, «Нажать левую») является многомодальной, а белый кружок обозначает одномодальную речевую команду (например, «Копировать» или «Вставить»).

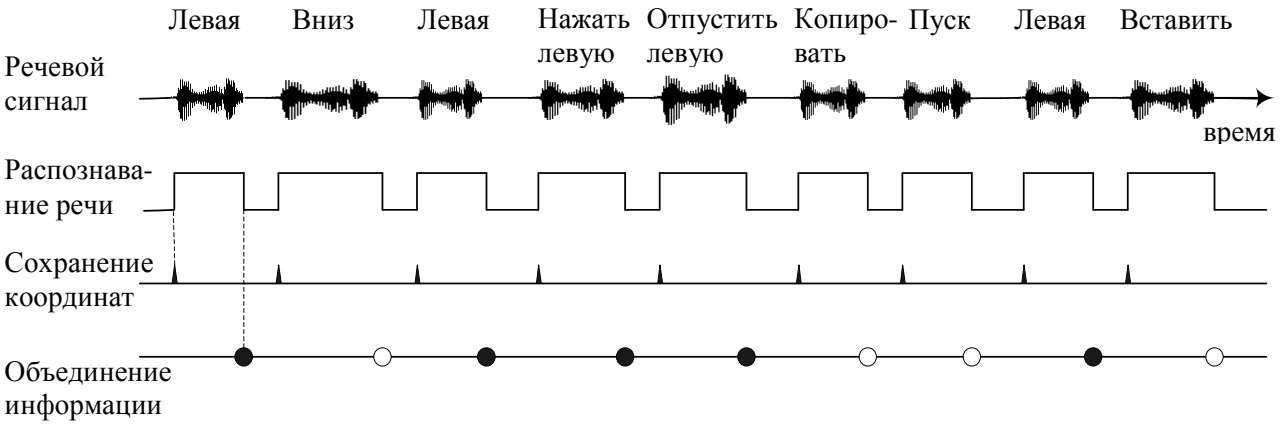

Рис. 4.12. Пример синхронизации потоков и объединения информации

Текущее положение курсора сохраняется в буфере системы в момент определения наличия речи оператора (срабатывания алгоритма поиска границ речи). Это связано с тем, что во время произнесения фразы пользователь может непреднамеренно перемещать голову и тем самым менять положение курсора, в результате чего он будет указывать на другой графический объект. Кроме того, речевое намерение формируется в сознании в соответствии с целью и ситуацией до того как произносится голосовая команда. После окончания процесса распознавания произнесенной пользователем команды модуль распознавания речи дает сигнал на объединение информации и выполнение многомодальной команды.

Для объединения информации, поступающей от двух модальностей, используется фреймовый метод позднего объединения, когда поля определенной структуры (фрейма) заполняются данными по мере их поступления, а по окончании процесса распознавания, выполняется многомодальная команда. В том случае если распознанная команда является многомодальной (см. таблицу 4.1), то она объединяется в один фрейм с сохраненными координатами курсора и автоматически посылается сообщение

виртуальному устройству мыши о выполнении нужного действия. Если же голосовая команда является одномодальной, то координаты курсора не учитываются и посылается соответствующее сообщение виртуальному устройству клавиатуры.

4.3.2. Модуль распознавания голосовых команд оператора

Звуковой сигнал, непрерывно записываемый микрофоном цифровой видеокамеры, обрабатывается в модуле распознавания речи. Процесс распознавания речи запускается детектором границ речи, который обнаруживает наличие речи в звуковом сигнале. Процесс распознавания заканчивается после получения наилучшей гипотезы распознавания голосовой команды. В таблице 4.1 представлен перечень основных голосовых команд, используемых для управления персональным компьютером без помощи рук. Приведенный набор содержит 40 голосовых команд, которые являются наиболее часто используемыми командами при работе с графическим пользовательским интерфейсом. Теоретически, возможно работать с компьютером, используя лишь левую и правую кнопки мыши (команды «Левая» и «Правая»), однако введение дополнительных голосовых команд позволяет значительно ускорить взаимодействие.

Все голосовые команды можно разделить на четыре класса по их функциональному назначению: команды замещающие работу мыши, команды замещающие работу клавиатуры, команды для графического пользовательского интерфейса, а также специальный класс, который содержит служебные команды (на данный момент используется лишь одна команда - «Калибровка», предназначенная для запуска процесса настройки системы отслеживания движений головы). Нужно отметить, что лишь команды замещающие работу мыши являются многомодальными (тип М), так как они используют информацию о положении курсора в текущий момент времени. Все остальные команды являются одномодальными (речевыми) командами (тип О) и при их выполнении положение курсора не учитывается.

Таблица 4.1. Перечень голосовых команд модели бесконтактного управления компьютером

Класс команд	Текст команды	Действие	Тип
команды замещающие работу мыши	левая	кликнуть левой кнопкой мыши	М
	правая	кликнуть правой кнопкой мыши	М
	нажать левую	зажать левую кнопку мыши	М
	отпустить левую	отпустить левую кнопку мыши	М
	нажать правую	зажать правую кнопку мыши	М
	отпустить правую	отпустить правую кнопку мыши	М
	двойной клик	дважды кликнуть левой кнопкой мыши	М
	вниз	прокрутка окна вниз (колесико мыши)	О
	вверх	прокрутка окна вверх (колесико мыши)	О
команды замещающие работу клавиатуры	ввод	нажать кнопку «Enter»	О
	0-9	нажать цифровую кнопку клавиатуры	О
	выйти	нажать кнопку «Escape»	О
	удалить	нажать кнопку «Delete»	О
	регистр	нажать кнопку «Caps Lock»	О
	выключить	нажать кнопку «Power» клавиатуры	О
команды графического пользова-тельского интерфейса	новый	создать пустой документ	О
	открыть	открыть файл	О
	сохранить	сохранить файл	О
	закрыть	закрыть активное окно	О
	копировать	копировать выделенный фрагмент	О
	вырезать	вырезать выделенный фрагмент	О
	вставить	вставить фрагмент из буфера	О
	печать	вывод файла на печать	О
	найти	диалоговое окно «Найти»	О
	отменить	отменить действие	О
	вперед	показать следующее окно	О
	назад	показать предыдущее окно	О
	выделить все	выделить всю страницу	О
	пуск	открыть меню «Пуск»	О
	помощь	открыть меню «Помощь»	О
специальные команды	калибровка	запуск процесса настройки модуля отслеживания головы	О

4.3.3. Эксперименты с моделью бесконтактной работы с компьютером

Для тестирования многомодальной модели были привлечены пять независимых тестеров, которые имели незначительный опыт работы с персональным компьютером, а также один пользователь с ограниченными возможностями (без рук), проходящий реабилитацию в Санкт-Петербургском профессионально-реабилитационном центре [58]. В ходе экспериментов пользователи должны были выполнить несколько сценариев работы с компьютером. Эти задачи включали в себя управление графическим редактором MS Paint, текстовым редактором MS Word, а также доступ к ресурсам сети Интернет посредством броузера MS Internet Explorer. В таблице 4.2 представлена последовательность действий, которую пользователи должны были выполнить, используя предложенную многомодальную модель ICanDo, а также обычным способом, используя клавиатуру и мышь. Данный тестовый сценарий включает в себя следующие действия: нахождение информации о программе телепередач (канала НТВ) на Интернет портале Рамблер «www.rambler.ru», скопировать фрагмент данной вэб-страницы в буфер, открыть текстовый редактор MS Word, вставить в пустой документ информацию из буфера, сохранить файл на рабочем столе компьютера и распечатать файл. Эту задачу можно разбить на цепочку элементарных действий, которые тестеры-пользователи выполняли при помощи разработанной многомодальной системы (речь + движения головы), а также традиционным способом (мышь + клавиатура).

Таблица 4.2. Ход выполнения тестового сценария

№	Описание действия	Способ взаимодействия	
		голова + речь	**мышь + клавиатура**
1	выделение ссылки *ТВ*	движение головой	движение мышкой
2	нажатие ссылки *ТВ*	команда «Левая»	клик левой кнопкой

3	прокрутка окна вниз	команда «Вниз»	колесо мыши на себя
4	прокрутка окна вниз	команда «Вниз»	колесо мыши на себя
5	выделение ссылки *HTB*	движение головой	движение мышкой
6	нажатие ссылки *HTB*	команда «Левая»	клик левой кнопкой
7	установка курсора	движение головой	движение мышкой
8	зажать левую кнопку	ком. «Нажать левую»	зажать левую кнопку
9	установка курсора	движение головой	движение мышкой
10	отпустить левую кнопку	команда «Отпустить левую»	отпустить левую кнопку
11	копировать выделенное	команда «Копировать»	нажать «Ctrl+C»
12	открыть меню *Start*	команда «Пуск»	движение мышкой
			клик левой кнопкой
13	выделение ярлыка *Word*	движение головой	движение мышкой
14	запустить *Word*	команда «Левая»	клик левой кнопкой
15	вставка из буфера	команда «Вставить»	нажать «Ctrl+V»
16	сохранение файла	команда «Сохранить»	нажать «Ctrl+S»
17	выделение пункта *Folder*	движение головой	движение мышкой
18	открытие дерева каталогов	команда «Левая»	клик левой кнопкой
19	выделение пункта *Desktop*	движение головой	движение мышкой
20	выбрать элемент *Desktop*	команда «Левая»	клик левой кнопкой
21	выделение кнопки *Save*	движение головой	движение мышкой
22	нажатие кнопки *Save*	команда «Левая»	клик левой кнопкой
23	открытие меню печати	команда «Печать»	нажать «Ctrl+P»
24	выделение кнопки *Print*	движение головой	движение мышкой
25	нажатие кнопки *Print*	команда «Левая»	клик левой кнопкой
26	закрытие *Word*	команда «Закрыть»	движение мышкой
			клик левой кнопкой
27	закрытие *Internet Explorer*	команда «Закрыть»	движение мышкой
			клик левой кнопкой

Таблица 4.3 показывает количественные результаты экспериментов и сравнение скорости работы (управления) двух способов взаимодействия (время, требуемое для выполнения тестового сценария каждым из пользователей и точность распознавания речи). Время выполнения сценария стандартным способом для пользователя 6 отсутствует, так как он является инвалидом и не может работать с мышкой и клавиатурой. Точность распознавания голосовых команд составила свыше 96% для каждого из пользователей. Следует отметить, что по мере обучения работе с многомодальной моделью пользователи увеличивали скорость своей работы. В то же время пользователи без опыта работы с компьютером выполняли сценарий многомодальным способом почти с такой же скоростью, как при использовании стандартных органов управления.

Таблица 4.3. Сравнение способов управления компьютером

Тестер	Точность распознавания голосовых команд, %	Время выполнения сценария, сек.	
		голова + речь	мышь + клавиатура
1	98	84	43
2	97	73	36
3	97	91	44
4	97	88	50
5	96	77	42
6	98	80	-
Средн.	97	82	43

Таким образом, в ходе проведенных экспериментов было установлено, что предложенный многомодальный способ управления оказался в 1.9 раз медленнее, чем клавиатурно-ориентированный способ. Однако, такое замедление приемлемо, так как модель разрабатывается для пользователей со специальными требованиями, в частности, для пользователей без рук или с парализованными руками, помогая им в процессе социально-экономической интеграции с информационным обществом и делая их более независимыми от помощи со стороны других лиц.

Выводы по главе 4

1) Представлена программная реализация SIRIUS, включающая в себя разработанные в ходе диссертационной работы модели обучения и распознавания русской речи, а также модули для записи речевых данных и анализа результатов распознавания речи.
2) Приведено сравнение реализованных целословной, морфемной, слоговой и фонемной моделей автоматического распознавания русской речи по критериям точности распознавания на уровне фонем, слов и фраз, а также времени распознавания фраз. Отмечен рост скорости распознавания при применении морфемной модели на 75% при незначительном падении точности распознавания.
3) Разработанные в диссертации модели были реализованы в диалоговой модели голосового доступа к электронному справочному каталогу, представлена архитектура данной модели и результаты ее тестового применения. Предложенная модель голосового доступа, не учитывает концовки слов в распознанной гипотезе фразы, что позволяет поднять точность распознавания фраз на 7% по сравнению с базовой морфемной моделью и практически достичь уровня целословного распознавания речи.
4) Реализована многомодальная модель ICanDo для бесконтактного управления компьютером (полностью без клавиатуры и мышки), использующая модули автоматического распознавания голосовых команд и отслеживания положения головы оператора для работы с графическим пользовательским интерфейсом персонального компьютера. Модель предназначается, в основном, для пользователей-инвалидов, имеющих проблемы с двигательными функциями рук и показала хорошие результаты в ходе экспериментов с потенциальными пользователями.

Заключение

При автоматическом распознавании речи для любого языка существует ряд общих проблем, для решения которых прикладываются усилия ученых и разработчиков всего мира. Основными из них являются: обеспечение слитного ввода речи, дикторонезависимость, повышение точности и скорость обработки речи, улучшение робастности систем и т.д. Вторую группу проблем составляют особенности конкретных языков, для которых применяются распознаватели. Так, для русского языка (а также для многих славянских и ряда других языков) такой особенностью является сложный механизм словообразования, из-за чего образуется огромное количество словоформ в языке.

В результате проведенной работы были разработаны модели, предназначенные для обработки русской слитной речи на различных уровнях и предложен подход к распознаванию, использующий морфемный уровень представления русского языка и речи, что позволяет значительно сократить количество распознаваемых языковых единиц в модели и за счет этого повысить скорость обработки, сохранив при этом приемлемую точность распознавания.

Предложены модели обучения распознавателя речи и дикторонезависимого распознавания русской слитной речи. Модель обучения позволяет создать статистическую морфемную модель языка прикладной области, морфемный словарь с фонетическими транскрипциями и акустико-фонетические модели русской речи. Эти базы данных необходимы для функционирования модели распознавания, которая содержит морфемный уровень представления русского языка и речи, и позволяет производить распознавание речи по морфемным единицам с последующим синтезом гипотез слов и фраз из цепочек морфем. В ходе экспериментов морфемная модель сравнивалась с целословной, слоговой и фонемной моделями и показала оптимальные результаты по критериям точности распознавания и времени обработки.

Предложенные модели обучения и распознавания объединены в программную реализацию SIRIUS, позволяющую производить дикторонезависимое распознавание русской слитной речи с большим словарем. На базе данной реализации разработаны прикладные модели для голосового доступа к справочному электронному каталогу и многомодальная модель для бесконтактного управления компьютером. Разработанные в диссертационной работе методы, модели и программные средства будут использованы в дальнейшем при создании интеллектуальных приложений человеко-машинного речевого и многомодального взаимодействия.

Литература

1. Александров, В.В. Структурный анализ диалога / В.В. Александров, А.В. Арсентьева, А.И. Семенков // Ленинград: ЛНИВЦ, 1983, 49 с.
2. Афанасьев, В.П. Архитектура речевого телефонного терминала МАРС-2 «Электроника МС7602» / В.П. Афанасьев и др. // Труды Всесоюзного семинара АРСО-14, Каунас , 1986, С. 77.
3. Баранников, В.А. Пакет программ построения систем распознавания речи / В.А. Баранников, А.А. Кибкало // Труды III Всероссийской конференции «Теория и практика речевых исследований» АРСО-2003. Москва, МГУ им. М.В. Ломоносова, Сентябрь 2003, С. 7-12.
4. Беллман, Р. Динамическое программирование / Р. Беллман; М.: ИЛ, 1960, 400 с.
5. Бияков, О.А. Медианное сглаживание временных рядов / О.А. Бияков // Вестник КузГТУ. 1999. № 3. С. 55 -56.
6. Васьков, С.Т. Открытые системы реального времени / С.Т. Васьков, В.Н. Вьюхин, И.И. Коршевер // Информатика и вычислительная техника. - М.: Изд-во ВИМИ, 1995, вып. 1-2, С. 96-106.
7. Винцюк, Т.К. Модуль анализатора речи СРД «Речь-2» / Т.К. Винцюк, А.Г. Скрипник // Тезисы докладов 16-го всесоюзного семинара (АРСО – 16), 1991. – С. 250-251.
8. Винцюк, Т. К. Распознавание слов устной речи методами динамического программирования/ Т. К. Винцюк // М.: Кибернетика, 1968. – №1. – С. 15-22.
9. Галунов, В.И. Состояние исследований в области речевых технологий и задачи, выдвигаемые государственными заказчиками / В.И. Галунов, и др. // Доклад на секции по автоматическому распознаванию и синтезу речи РАН. М., 2002.
10. Галушкин, А.И. Теория нейронных сетей / А.И. Галушкин; М.:ИПРЖР, 2000, 416 с.

11. Геппенер, В.В. Вейвлет-преобразование в задачах цифровой обработки сигналов: Учебное пособие / В.В. Геппенер, Д.А. Черниченко, С.А. Экало // СПб.: Изд-во СПбГЭТУ, 2002. 78 с.
12. Гринберг, Д. Квантитативный подход к морфологической типологии языков / Д. Гринберг // НЛ. Вып. III. М., 1963.
13. Дегтярев, Н.П. Параметрическое и информационное описание речевых сигналов / Н.П. Дегтярев // Минск: Объединенный институт проблем информатики НАН Беларуси, 2003, 216 с.
14. Джелинек, Ф. Разработка экспериментального устройства, распознающего раздельно произносимые слова / Ф. Джелинек // ТИИЭР. Речевая связь с машинами. т.73, №11, 1985, с. 91-100.
15. Джелинек, Ф. Распознавание непрерывной речи статистическими методами / Ф. Джелинек // ТИИЭР 64, № 4, 1976, с. 131-160.
16. Зализняк, А.А. Грамматический словарь русского языка: Словоизменение / А.А. Зализняк // 4-е изд., испр. и доп. — М.: Руские словари, 2003.
17. Иванова, Т.И. Компьютерные технологии в телефонии / Т.И. Иванова // Эко-Трендз, М., 2002.
18. Карпов, А.А. SIRIUS - система дикторонезависимого распознавания слитной русской речи / А.А. Карпов, А.Л. Ронжин, И.В. Ли // Известия ТРТУ, № 10, 2005, С. 44-53.
19. Карпов, А.А. Система акустических признаков речевого сигнала, устойчивых к вариациям уровня громкости и спектра сигнала / А.А. Карпов и др. // Труды 3-й Всероссийской конференции «Теория и практика речевых исследований» АРСО’2003, Москва, 2003, с. 83-88.
20. Кибкало, А.А. Разработка системы распознавания русской речи / А.А.Кибкало и др. // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2003. Вып. 3. С. 8-20.
21. Косарев, Ю.А. Естественная форма диалога с ЭВМ / Ю.А. Косарев – Л.: Машиностроение, 1989. – 143 с.

22. Ли, И.В. Проектирование систем речевого диалога / И.В. Ли, А.Л. Ронжин // Труды СПИИРАН. Вып. 3, т. 1. — СПб.: Наука, 2006, С. 320-338.
23. Мазуренко, И.Л. Многоканальная система распознавания речи / И.Л. Мазуренко // Сборник трудов VI всероссийской конференции «Нейрокомпьютеры и их применение», Москва, 2000.
24. Маркел, Д.Д. Линейное предсказание речи / Д.Д. Маркел, А.Х. Грей; М.: Связь, 1980. – 308 с.
25. Марков, А.А. Об одном применении статистического метода / А.А. Марков // Известия АН, сер.6, X, №4, 1916, 239 с.
26. Моттль, В.В. Скрытые Марковские модели в структурном анализе сигналов / В.В. Моттль, И.Б. Мучник; М.: Физматлит, 1999, 351 с.
27. Мясников, Л.Л. Объективное распознавание звуков речи / Л.Л. Мясников // ЖТФ. – 1943. –№ 3. – С. 109-115.
28. Никифоров, В.О. Адаптивное и робастное управление с компенсацией возмущений / В.О. Никифоров // СПб.: Наука, 2003, 282 с.
29. Осовский, С. Нейронные сети для обработки информации / С. Оссовский, пер. с польского И. Рудинского. -М.: Финансы и статистика, 2004, 344 с.
30. Петровский, А.А. Методы построения устройств распознавания речи на базе гибрида нейронная сеть/скрытая Марковская модель / А.А. Петровский // Нейрокомпьютеры: разработка, применение, 2002, № 12, с. 26-36.
31. Потапова, Р.К. Речевое управление роботом / Р.К. Потапова // М.:КомКнига, 2005, 328 с.
32. Потапова, Р.К. Речь: коммуникация, информация, кибернетика / Р.К. Потапова // М.:Едиториал УРСС, 2003, 568 с.
33. Рабинер, Л. СММ и их применение в избранных приложениях при распознавании речи / Л. Рабинер // ТИИЭР. – 1989. – Т. 77. – №2. – С. 86-120.
34. Рабинер, Л. Цифровая обработка речевых сигналов / Л. Рабинер, Р. Шафер – М.: Радио и связь, 1987.

35. Распознавание слуховых образов. / Под ред. Н.Г. Загоруйко – Новосибирск: «Наука», 1970. – 340 с.
36. Ронжин, А. Метод распознавания слитной речи на основе анализа сигнала в скользящем окне и теории размытых множеств / А. Ронжин и др. // Научно-теоретический журнал «Искусственный интеллект», №4. – Донецк, Украина, 2002, С. 256-263.
37. Ронжин, А.Л. Речевой и многомодальный интерфейсы / А.Л. Ронжин, А.А. Карпов, И.В. Ли; - М.: Наука, 2006 - (Информатика: неограниченные возможности и возможные ограничения), 173 с.
38. Ронжин, А.Л. Система автоматического распознавания русской речи SIRIUS / А.Л. Ронжин, А.А. Карпов, И.В. Ли // Научно-теоретический журнал «Искусственный интеллект», № 3. – Донецк, Украина, 2005, С. 590-601.
39. Ронжин, А.Л. Фонетико-морфологическая разметка речевых корпусов для распознавания и синтеза русской речи / А.Л. Ронжин и др. // Информационно-управляющие системы, Вып. 25, т. 6. — СПб.: ГУАП, 2006, С. 24-34.
40. Русская грамматика: [В 2 т. / Редкол.: Н.Ю. Шведова (гл. ред.) и др.]. Т. 1: Фонетика. Фонология. Ударение. Интонации. Словообразование. Морфология / [Н.С. Авилова, А.В. Бондарко, Е.А. Брызгунова и др.] - М. : Наука, 1980, 783 с.
41. Сайт белорусской компании Сакрамент http://www.sakrament.com/viewprod.php?TopId=30&ProdId=24
42. Сайт библиотеки М. Мошкова http://www.lib.ru/
43. Сайт инструментария Hidden Markov Model Toolkit http://htk.eng.cam.ac.uk/
44. Сайт компании Истрасофт http://www.istrasoft.ru/voice_cmd.html
45. Сайт телекоммуникационной компании NewVoice http://www.newvoice.ru/
46. Сайт компании Nuance Corporation http://www.nuance.com
47. Сайт конкурса многомодальных интерфейсов Loco Mummy Contest http://www.locomummy.net

48. Сайт проекта Oomnik / Корнеслов http://www.oomnik.ru
49. Сайт Европейского проекта FP6 SIMILAR Network of Excellence http://www.similar.cc
50. Сайт проекта STARLING http://starling.rinet.ru
51. Сайт рабочей группы АОТ http://www.aot.ru
52. Сайт системы «Telepat» https://www.telepat.ru
53. Сайт системы «Горыныч» http://www.nd.ru/voice/
54. Сайт системы SAMPA http://www.phon.ucl.ac.uk/home/sampa/home.htm
55. Сайт электронного каталога «Желтые страницы» http://www.yell.ru/
56. Сайт Центра речевых технологий http://speechpro.com/production/?id=471&fid=44
57. Сапожков, М.А. Речевой сигнал в кибернетике и связи / М.А. Сапожков; – М.: Связьиздат, 1963. – 452 с.
58. Сборник статей профессионально-реабилитационного центра Санкт-Петербурга, «Человек и здоровье», 2006, 135 с.
59. Скрелин, П.А. Сегментация и транскрипция / П.А. Скрелин; СПб.: СПбГУ, 1999.
60. Современный русский литературный язык / под ред. П.А. Леканта. М., 1996, 160 с.
61. Сокирко, А.В. Морфологические модули на сайте www.aot.ru / А.В. Сокирко // Труды Международной конференции Диалог-2004, М.: Наука, 2004. С. 559.
62. Соколов, Б.В. Концептуальные основы оценивания и анализа качества моделей и полимодельных комплексов / Б.В. Соколов, Р.М. Юсупов // Теория и системы управления. – 2004. –№ 6 –С. 5–16.
63. Станкевич, Л.А. Интеллектуальные роботы и системы управления / Л.А. Станкевич // Нейрокомпьютеры: разработка и применение, № 8-9, 2005.
64. Страуструп, Б. Язык программирования С++ / Б. Страуструп // М.:БИНОМ, 2001. – 1099 с.

65. Трунин-Донской, В.Н. Опознавание набора слов с помощью цифровой вычислительной машины / В.Н. Трунин-Донской // Работы по технической кибернетике. – М.: ВЦ АН СССР, 1967. – С. 37-51.
66. Ушакова, Т.Н. Проблема внутренней речи в психологии и психофизиологии. Психологические и психофизиологические исследования речи / Т.Н. Ушакова – М.: Наука, 1985. – С. 13-26.
67. Фант, Г. Анализ и синтез речи / Г. Фант; пер. с англ. В.С. Лозовского и Н.В. Бахмутовой под ред. Н.Г. Загоруйко. Новосибирск, «Наука», 1970,167 с.
68. Холоденко, А.Б. Использование лексических и синтаксических анализаторов в задачах распознавания для естественных языков / А.Б. Холоденко // Интеллектуальные системы. Т.4, вып. 1-2, 1999, с. 185-193.
69. Холоденко, А.Б. О построении статистических языковых моделей для систем распознавания русской речи / А.Б. Холоденко // Интеллектуальные системы. т.6, вып. 1-4, 2002. С. 381-394.
70. Цымбал, В.П. Теория информации и кодирование / В.П. Цымбал // Киев.:Высшая Школа, 1977, 288 с.
71. Чучупал, В.Я. К вопросу об оптимальном выборе алфавита моделей звуков русской речи для распознавания речи / В.Я. Чучупал, К.А. Маковкин, А.В. Чичагов // Искусственный интеллект, 2002, №2, с. 575-579.
72. Шелепов, В.Ю. К проблеме пофонемного распознавания / В.Ю. Шелепов, В.Ю. Ниценко // Искусственный интеллект. - 2005. - № 4. - С. 662-668.
73. Щерба, Л.В. Языковая система и речевая деятельность / Л.В. Щерба; Л., 1974.
74. Arisoy, E. A Unified Language Model for Large Vocabulary Continuous Speech Recognition of Turkish / E. Arisoy, et al // Signal Processing, № 86(10), 2006, pp. 2844-2862.

75. Atal, B.S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification / B.S. Atal // Journal of the Acoustical Society of America, Vol. 55, 1974, pp. 1304-1312.
76. Baum, L.E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes / L.E. Baum // Inequalities, vol.3, 1972, pp. 1-8.
77. Bouguet, J.-Y. Pyramidal implementation of the Lucas-Kanade feature tracker / J.-Y. Bouguet // Technical Report, Intel Corporation, Microprocessor Research Labs, 2000.
78. Chesta, C. Connected Digit Recognition Using Short and Long Duration Models / C. Chesta, P. Laface, F. Ravera // Proceedings of ICASSP'99 Conference, Phoenix, USA, 1999.
79. Cox, R.V. Speech and Language Processing for Next-Millennium Communications Services / R.V. Cox, et al // Proceedings of the IEEE, Vol. 88, No. 8, 2000, pp. 1314-1337.
80. Creutz, M. Unsupervised discovery of morphemes / M. Creutz, K. Lagus. // Proceedings of ACL/SIGPHON'2002, 2002, pp. 21–30.
81. Freeman, D. A Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service / D. Freeman, C. Sonthcott, I. Boyd // IEEE Colloquium Digitized Speech Communication via Mobile Radio, 1988, pp. 61-65.
82. Fujimoto, M. Evaluation of noisy speech recognition based on noise reduction and acoustic model adaptation on the AURORA2 tasks / M. Fujimoto, Y. Ariki // Proceedings of ICSLP'2002, Denver, USA, 2002.
83. Furui, S. 50 years of progress in speech and speaker recognition / S. Furui // Proceedings of SPECOM'2005, Patras, Greece, 2005, pp. 3-9.
84. Haton, J.-P. Automatic speech recognition: Past, Present and Future / J.-P. Haton // Proceedings of SPECOM'2004, St. Petersburg: "Anatoliya", 2004, pp. 3-7.

85. Hirsimaki, T. Unlimited Vocabulary Speech Recognition with Morph Language Models Applied to Finnish / T. Hirsimaki, et al // Computer Speech and Language, Vol. 20, № 4, 2006, pp. 515-541.
86. Hori, T. An extremely-large-vocabulary approach to named entity extraction from speech / T. Hori, A. Nakamura // Proceedings of ICASSP'2006, Toulouse, France, 2006.
87. International Phonetic Association. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge: Cambridge University Press, 1999.
88. Jelinek, F. Perplexity - A measure of difficulty of speech recognition tasks / F. Jelinek, R.L. Mercer, L.R. Bahl // Proceedings of 94-th Meeting of the Acoustical Society of America, 1977.
89. Kanevsky, D., Monkowski M., Sedivy J. Large vocabulary speaker-independent continuous speech recognition in Russian language / D. Kanevsky, M. Monkowski, J. Sedivy // Proceedings of SPECOM'1996, St.Petersburg, 1996, pp.117-121.
90. Kanungo, T. An Efficient k-Means Clustering Algorithm: Analysis and Implementation / T. Kanungo, et al // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002, vol. 24, №7. p. 881-892.
91. Karpov, A. A multi-modal system ICANDO: Intellectual Computer AssistaNt for Disabled Operators / A. Karpov, A. Ronzhin, A. Cadiou // Proceedings of Interspeech'2006, Pittsburgh, PA, USA, 2006, pp. 1998-2001.
92. Karpov, A. Hands-free Mouse Control System for Handicapped Operators / A. Karpov, A. Cadiou // Proceedings of SPECOM'2006, St. Petersburg: "Anatoliya", 2006, pp. 525-529.
93. Katz, S. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer / S. Katz // IEEE Transactions on Acoustic, Speech and Signal Processing, 1987, vol. 35, №. 3, pp. 400-401.
94. Kosarev, Yu. Robust Speech Understanding for a Voice Control System / Yu. Kosarev, et al // Proceedings of SPECOM'2002, St. Petersburg, 2002, pp. 13-18.

95. Kurimo, M. Unsupervised Segmentation of Words into Morphemes - Morpho Challenge 2005. Application to Automatic Speech Recognition / M. Kurimo, et al // Proceedings of Interspeech'2006, Pittsburg, USA, pp. 1021-1024.
96. Kwon, O.W. Korean large vocabulary continuous speech recognition with morpheme-based recognition units / O.W. Kwon, J. Park // Speech Communication, №39, 2003, pp. 287-300.
97. Manning, C.D. Foundations of Statistical Natural Language Processing / C.D. Manning, H. Schütze; MIT Press, 1999.
98. Oparin, I. Stem-Based Approach to Pronunciation Vocabulary Construction and Language Modeling for Russian / I. Oparin, A. Talanov // Proceedings of SPECOM'2005, Patras, Greece, 2005, pp. 575-578.
99. Oviatt, S.L. Multimodal interfaces / S.L. Oviatt // Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. Lawrence Erlbaum Assoc. Mahwah, NJ, USA, 2003, pp. 286-304.
100. Picone, J. Continuous Speech Recognition Using Hidden Markov Models / J. Picone // IEEE ASSP Magazine, Vol. 7, No. 3, 1990.
101. Pollard, C.J. Head-driven Phrase Structure Grammar / C.J. Pollard, I.A. Sag; Chicago University Press, Chicago, 1994.
102. Potapova, R. Identification of prosodic features of emotional state of a speaker / R. Potapova, V. Potapov // Proceedings of SPECOM'2005. Patras, Greece, 2005, pp. 25-32.
103. Potapova, R. To the problem of multi-language phonetic database formation: vibrants in English, German, Russian and Chechen / R. Potapova, E. Loseva // Proceedings of SPECOM'2006, St. Petersburg: "Anatoliya", 2006, pp. 445-448.
104. Potryasaev, S. Quality and Quantity Estimation and Analysis of Multimodal Systems for Human-Computer Interaction / S. Potryasaev, B. Sokolov, R. Yusupov // Proceedings of SPECOM'2006, St. Petersburg: "Anatoliya", 2006, pp. 158-167.

105. Psutka, J. Large Vocabulary ASR for Spontaneous Czech in the MALACH Project / J. Psutka, et al // Proceedings of Eurospeech'2003, Geneva, Switzerland, 2003, pp. 1821-1824.
106. Rabiner, L. Fundamentals of Speech Recognition / L. Rabiner, B. Juang – New Jersey: Prentice-Hall, Englewood Cliffs, USA, 1993.
107. Rabiner, L.R. A tutorial on Hidden Markov Models and Selected Applications in SpeechRecognition / L.R. Rabiner // Proceedings of the IEEE, vol, 77. no.2, 1989, pp. 257-284.
108. Shen, J.-L. Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments / J.-L. Shen, J.-W. Hung, L.-S. Lee // Proceedings of ICSLP'1998, Sydney, Australia, 1998.
109. Strom, N. Continuous Speech Recognition in the WAXHOLM Dialogue System / N. Strom // Stockholm QPSR, 1996. – pp. 67-95.
110. Surendran, D. Dialog Act Tagging with Support Vector Machines and Hidden Markov Models / D. Surendran, G. Levow // Proceedings of Interspeech'2006, Pittsburgh, PA, USA, 2006, pp. 1950-1953.
111. Tang, M. Improvements to Bucket Box Intersection Algorithm for Fast GMM Computation in Embedded Speech Recognition Systems / M. Tang, A. Ganapathiraju // Proceedings of Interspeech'2006. Pittsburgh, USA, pp. 617-620.
112. Timofeev, A.V. Development of man-machine interfaces and virtual reality means for integrated medical systems / A.V. Timofeev, et al. // Proceedings of SPECOM'2006, St. Petersburg: "Anatolya", 2006, pp. 175-178.
113. Trentin, E. A survey of hybrid ann/hmm models for automatic speech recognition / E. Trentin, M. Gori // Neurocomputing, vol. 37, no. 1-4, 2001, pp. 91-126.
114. Turunen, M. Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences / M. Turunen, J. Hakulinen, A. Kainulainen // Proceedings of Interspeech'2006, Pittsburgh, USA, 2006, pp. 1057-1060.

115. Varile, G. Survey of the State of the Art in Human Language Technology / G. Varile, A. Zampolli // Cambridge University Press, 1997.
116. Viterbi, A.J. Error bounds for convolutionalcodes and an asymmetrically optimum decoding algorithm / A.J. Viterbi // IEEE Transactions on Information Theory, 1967. – vol. IT-13, pp. 260-267.
117. Waheed, K. A robust algorithm for detecting speech segments using an entropy contrast / K. Waheed, K. Weaver, F. Salam // Proceedings of MWSCAS'2002, Oklahoma, USA, 2002.
118. Whittaker, E.W.D. Statistical Language Modelling for Automatic Speech Recognition of Russian and English / E.W.D. Whittaker // PhD thesis, Cambridge University, Cambridge, 2000.
119. Young, S. The HTK Book / S. Young, et al // Cambridge University Engineering Department, 2002.
120. Young, S.J. Token Passing: A Conceptual Model for Connected Speech Recognition Systems / S.J. Young, N.H. Russel, J.H.S. Russel // CUED Technical Report, Cambridge University, 1989.
121. Zhozhikashvili, V.A. The First Voice Recognition Applications in Russian Language for use in the Interactive Information Systems / V.A. Zhozhikashvili, et al // Proceedings of SPECOM'2004, St. Petersburg: "Anatoliya", 2004, pp. 304-308.