На правах рукописи

Чистиков Павел Геннадьевич

МЕТОДЫ И АЛГОРИТМЫ ГИБРИДНОГО СИНТЕЗА ЕСТЕСТВЕННОЙ РУССКОЙ РЕЧИ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ И МЕТОДА UNIT SELECTION

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2013

Работа выполнена в Санкт-Петербургском национальном исследовательском университете информационных технологий, механики и оптики (НИУ ИТМО) на кафедре речевых информационных систем.

Научный руководитель:

Матвеев Юрий Николаевич,
доктор технических наук, профессор кафедры речевых информационных систем НИУ ИТМО.

Официальные оппоненты:

Лобанов Борис Мефодьевич,
доктор технических наук, главный научный сотрудник лаборатории распознавания и синтеза речи Объединенного института проблем информатики НАН Беларуси,

Тропченко Андрей Александрович,
кандидат технических наук, доцент, доцент кафедры вычислительной техники НИУ ИТМО.

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук.

Защита диссертации состоится «23» мая 2013 года в 17 часов 00 минут на заседании диссертационного совета Д.212.227.06 при НИУ ИТМО по адресу: 197101, г. Санкт-Петербург, Кронверкский пр., д. 49, конференц-зал центра интернет-образования.

С диссертацией можно ознакомиться в библиотеке НИУ ИТМО.

Автореферат разослан «22» апреля 2013 г.

Ученый секретарь
диссертационного совета Д.212.227.06 И. С. Лобанов

Общая характеристика работы

Актуальность темы. С развитием технологий автоматического синтеза речи, синтезированная речь становится все более и более естественной, приближенной к речи человека. Однако системы синтеза речи в современных человеко-машинных интерфейсах, системах виртуальной реальности и мультимедийного общения, по-прежнему обладают рядом недостатков, которые утомляют слушателя, не давая ощущения того, что с ними общается живой человек. Для уменьшения количества дефектов, присущих синтезированной речи, различными научными коллективами разрабатываются методы, позволяющие повысить естественность речи. В России наиболее заметные результаты в области автоматического синтеза речи получены в Санкт-Петербургском государственном университете (П.А. Скрелин, В.И. Галунов), Институте проблем передачи информации РАН (В.Н. Сорокин), Московском государственном лингвистическом университете (Р.К. Потапова), МГУ им. М. В. Ломоносова (О.Ф. Кривнова). Из стран СНГ наиболее значимые результаты получены в Объединенном институте проблем информатики Национальной академии наук Беларуси (Б.М. Лобанов). В данном исследовании произведена разработка программного средства преобразования текста в речь, объединяющего подходы к синтезу речи, основанные на скрытых марковских моделях и методе Unit Selection. Такое программное средство обеспечивает обратную связь человека с вычислительной машиной посредством речевого интерфейса.

Разработанная автором гибридная система синтеза речи обеспечивает «чтение» произвольного русского текста без специальной предварительной разметки, с максимальной приближенностью к естественной слитной речи и естественным тембром голоса в широком диапазоне изменения основного тона голоса диктора и темпа его речи. Такая система востребована во всех случаях, когда получателем информации является человек: разгружается зрительный аппарат и повышается интерактивность взаимодействия с компьютером. Особенно остро данная система необходима для людей с ограниченными возможностями, в частности, инвалидов по зрению.

Наряду с системой распознавания речи, система синтеза речи может быть использована в call-центрах и системах автоматического информирования. Приложения на его основе могут быть востребованы во всех информационных сервисах в случаях, когда необходимо осуществление коммуникационных действий с пользователем, а предварительная запись требуемых фраз по тем или иным причинам невозможна.

Актуальность проведенных исследований подтверждается большим количеством докладов на эту тему на международных научно-технических конференциях, крупнейшей из которых является ежегодная конференция

Interspeech, и потребностью рынка в программно-технических средствах, позволяющих осуществлять интерактивное взаимодействие с компьютером посредством речи.

В результате работы создано программное средство, обеспечивающее человеко-машинный интерфейс, где ЭВМ выполняет взаимодействие с человеком посредством голоса. Затронуты такие аспекты, как теоретическое и экспериментальное исследование в области систем управления базами данных и знаний (подготовка речевой базы данных и автоматизация этого процесса); разработка математического и программного обеспечения вычислительных машин (программные средства создания модели голоса и модификации речевого сигнала); повышение эффективности подготовки речевого корпуса (размеченной речевой базы данных) за счет автоматизации трудоемких процессов.

Цель диссертационной работы – создание программных средств синтеза естественной русской речи на основе совместного использования скрытых марковских моделей (СММ) и метода Unit Selection.

Для достижения данной цели были поставлены и решены **следующие задачи**.

1. Разработка методов, алгоритмов и программных средств синтеза естественной русской речи, основанных на совместном использовании скрытых марковских моделей и метода Unit Selection.
2. Проведение экспериментальных исследований, оценка качества работы созданной системы синтеза естественной русской речи, сравнение с мировыми аналогами.

Объектом исследования в данной работе являются системы преобразования печатного текста в естественно звучащую речь.

Предметом исследования является гибридная система синтеза естественной русской речи на основе совместного использования скрытых марковских моделей (СММ) и метода Unit Selection.

Научная новизна.

1. Разработана методика создания нового голоса для системы синтеза естественной русской речи, позволяющая существенно повысить качество звучания и снизить трудоемкость подготовки звуковой базы данных.
2. Разработаны алгоритмы стыковки и модификации речевых элементов, качественно улучшающие естественность синтезируемой речи.
3. Создан комплекс программных средств синтеза естественной русской речи на основе гибридной технологии, включающей совместное использование скрытых марковских моделей и метода Unit Selection.

Основные положения, выносимые на защиту.

1. Методика подготовки размеченной речевой базы данных (речевого корпуса).
2. Набор признаков звуковой единицы русского языка, обучение на основе которых приводит к созданию модели интонации, близкой к естественной.
3. Набор критериев поиска последовательности звуковых элементов методом Unit Selection, обеспечивающий высокое качество синтезированной речи.
4. Методика создания модели голоса.
5. Параллельные алгоритмы обучения моделей.
6. Алгоритм модификации частоты основного тона, энергии и длительности аллофонов.
7. Алгоритм стыковки звуковых элементов.

Методы исследования. В работе использованы методы дискретной математики, теории вероятностей и математической статистики, цифровой обработки сигналов, теории алгоритмов и прикладной лингвистики.

Достоверность научных положений, выводов и практических рекомендаций, полученных в диссертационной работе, подтверждается корректным обоснованием постановок задач, точной формулировкой критериев, компьютерным моделированием, результатами экспертной оценки, а также их внедрением на практике.

Практическая ценность. Результаты, полученные в ходе выполнения работы, используются на практике:

1) как самостоятельные решения, применяемые для озвучивания электронных книг и новостных лент;
2) в составе комплексного продукта, представляющего собой систему голосового самообслуживания.

Внедрение результатов работы. Результаты диссертации использованы при выполнении следующих научно-исследовательских работ: «Разработка комплекса аппаратно-программных средств синтеза русской речи по тексту» (федеральная целевая программа «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2012 годы»), «Разработка и реализация в виде программного обеспечения технологии синтеза речи на русском языке с учетом синтаксического и семантического анализа русского текста с высоким качеством звучания» и «Создание компьютерного лингвистического тренажера для экспресс-освоения навыков общения на иностранном языке» (по заказу министерства образования и науки РФ). Также результаты работы были внедрены в различные коммерческие продукты компании ООО «ЦРТ».

Апробация результатов работы. Основные положения диссертационной работы докладывались на научно-методических конференциях: «Международная конференция по компьютерной лингвистике Диалог-2010» (Москва), «Международная конференция по компьютерной лингвистике Диалог-2011» (Москва), «IEEE Conference, North West Russia Section» (Санкт-Петербург, 2011), «International Conference on Speech and Computer SPECOM 2011» (Казань), «XLI научная и учебно-методическая конференция НИУ ИТМО» (Санкт-Петербург, 2012), «I всероссийский конгресс молодых ученых НИУ ИТМО» (Санкт-Петербург, 2012), «Международная конференция по компьютерной лингвистике Диалог-2012» (Москва).

Личный вклад автора. Автором лично были разработаны программные средства синтеза русской речи на основе гибридной технологии, методика создания модели голоса и инструменты для ее обучения, алгоритмы модификации и стыковки звуковых элементов, качественно улучшающие естественность синтезируемой речи; проведены экспериментальные исследования по выбору признаков звуковых единиц русского языка и критериев поиска последовательности звуковых элементов методом Unit Selection. Реализована система сбора речевого материала, разметки, создания голоса синтеза. Подготовка основных публикаций проводилась с соавторами, при этом вклад автора был основным.

Публикации. По теме диссертации опубликовано 17 научных работ, в том числе 16 статей, из которых 6 статей опубликованы в журналах из перечня ВАК.

Структура диссертации. Диссертация изложена на 134-х страницах и состоит из введения, четырех глав и заключения. Список литературы содержит 132 наименования. Работа иллюстрирована 40-а рисунками и 13-ю таблицами.

Содержание работы

Во введении описывается предмет исследования, ставятся цель и задачи, обосновывается актуальность темы диссертационной работы. Формулируются положения, выносимые на защиту.

В первой главе приведен обзор современных методов и подходов к построению систем синтеза речи. Синтез речи представляет собой автоматическую генерацию речи на основе произвольного текста путем перевода последовательности символов данного текста в последовательность чисел, представляющих собой отсчеты звукового сигнала. Процесс построения системы синтеза речи можно разделить на несколько этапов (рисунок 1).

Синтез естественной (эмоциональной) речи возможен только на основе аллофонно и интонационно сбалансированного речевого корпуса. Такие корпуса имеют большой объем (не менее 10 часов речи), а их подготовка и

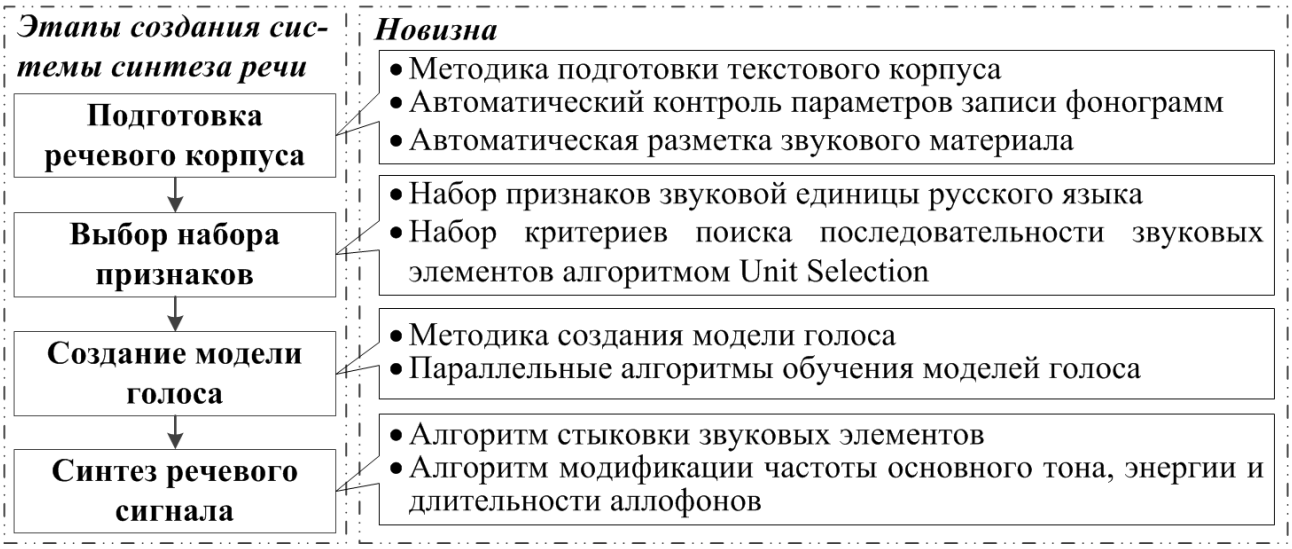

Рис. 1: Этапы создания системы синтеза речи, новизна

разметка осуществляется, как правило, вручную, что требует больших трудозатрат.

Также немаловажным является построение набора признаков звуковой единицы и разработка модели голоса, на основе которых будет формироваться синтезированная речь с присущими человеку интонациями и эмоциональной окраской (описание данных шагов представлено в последующих разделах).

Для генерации речевого сигнала существует несколько различных технологий, которые можно разделить на две группы: синтез, основанный на правилах, и синтез, основанный на данных. Первая технология была предложена еще в 1964 году и, несмотря на относительно плохое качество звучания, применяется по настоящее время. Вторая технология, появившаяся в 1977 году, позволяет добиться более высокого качества звучания.

Сравнивая подходы, основанные на методе Unit Selection и на статистических моделях, как наиболее распространенные, можно выделить следующие особенности.

1. Обучение параметров статистических моделей можно выполнять на относительно небольшом речевом материале, что позволяет существенно сократить объем требуемой для хранения речевого корпуса памяти, а также позволяет разрабатывать новый голос за гораздо меньший период времени за счет сокращения работ по разметке.
2. Речь, полученная на основе моделей, имеет более искусственное звучание, однако в ней отсутствуют разрывы, присутствующие при конкатенативном синтезе. При применении метода Unit Selection качество синтеза существенно ухудшается в случае отсутствия подходящего звукового элемента в базе данных. При применении моделей отсутствую-

щие в обучающей выборке звуковые элементы синтезируются на основе средних значений, максимально приближенных к требуемым, благодаря применению технологии кластеризации контекстов, основанной на деревьях. Это позволяет добиться разборчивости при ограниченном количестве контекстов.

3. Синтез, основанный на моделях, позволяет легко модифицировать характеристики голоса путем применения адаптации/интерполяции диктора, в то время как метод Unit Selection порождает речь, стиль которой не может быть отличен от стиля, представленного в речевом корпусе.

На основе вышеизложенного именно эти два подхода были выбраны для создания гибридной системы, как наиболее гармонично сочетающиеся и компенсирующие недостатки друг друга.

Во второй главе представлена методика создания нового голоса для системы синтеза речи; исследуются подходы, составляющие гибридную систему синтеза речи: на основе скрытых марковских моделей и методе Unit Selection; выбираются наборы признаков звуковой единицы русского языка и набор критериев поиска последовательности звуковых элементов, обеспечивающих высокое качество синтезируемой речи.

Процесс создания нового голоса для системы синтеза речи включает в себя три этапа: подготовка текстов для озвучивания человеком (диктором) (это должны быть сбалансированные интонационно представительные тексты), озвучивание подготовленных текстов (необходимо, чтобы диктор не допускал ошибок), разметка записанного звукового материала (является по большей части ручным трудоемким процессом). Автором диссертационной работы было разработано программное средство [5,9], полностью автоматизирующее этот процесс, позволяя автоматически формировать сбалансированный текст заданного объема для озвучивания диктором на основе статистической представительности звуковых элементов[17], выполнять контроль правильности чтения диктором заданных предложений, производить автоматическую разметку материала. Это позволило сократить время подготовки речевого корпуса до 20 раз, приблизив его к времени озвучивания текстового материала.

Система синтеза речи на основе скрытых марковских моделей состоит из двух основных частей:

1) обучение CMM-моделей, которые описывают параметры, полученные из обучающего звукового корпуса, учитывая контекстно-зависимые факторы;
2) составление объединенной CMM-модели, соответствующей синтезируемому тексту, и оценка наиболее вероятных параметров, на основе которых происходит синтез требуемого текста.

На этапе обучения моделей спектральные параметры, параметры функции возбуждения голосового тракта и информация о длительности состояний извлекаются из звуковой базы данных. Далее эти параметры описываются контекстно-зависимыми CMM-моделями для каждого аллофона (речевого звука). Определение контекста аллофона для конкретного языка является неотъемлемой составляющей, напрямую влияющей на качество результатов. Так, для русского языка, контекст был определен следующими признаками:

- аллофонные признаки: имена аллофона, стоящего перед предыдущим, предыдущего, текущего, следующего и следующего за следующим; позиция от начала и конца слога;
- слоговые признаки: типы предыдущего, текущего и следующего слогов; количество аллофонов в предыдущем, текущем и следующем слогах; позиция текущего слога в слове и синтагме; количество ударных слогов до и после текущего слога; тип гласной в слоге;
- словные признаки: часть речи предыдущего, текущего и следующего слов; количество слогов в предыдущем, текущем и следующем словах; позиция текущего слова в синтагме;
- синтагматические признаки: количество слогов и слов в синтагме; тип интонационного контура.

Суть метода Unit Selection заключается в выборе оптимальной последовательности звуковых элементов $u_1, u_2, ..., u_n$ на основе двух критериев: максимальной близости параметров выбранных элементов к требуемым параметрам и максимальной близости параметров на границах соседних звуковых элементов с целью обеспечения гладкости переходов. В качестве критерия выбора элементов используется стоимостная функция вида:

$$C(u,t) = \sum_{i=1}^{n} C^t(u_i, t_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i), \tag{1}$$

где $C^t(u_i, t_i)$ – стоимость замены, определяющая степень близости элемента u_i, взятого из базы данных, к требуемому элементу t_i, параметры которого определены на основе модели голоса; $C^c(u_{i-1}, u_i)$ – стоимость связи элементов u_{i-1} и u_i, определяющая степень гладкости перехода на границе данных элементов. В качестве оптимальных выбираются те элементы, значение функции (1) для которых минимально. Функция стоимости замены определяется следующим выражением:

$$C^t(u_i, t_i) = \sum_{k=1}^{p} \left(w_k^t C_k^t \left(u_i, t_i \right) \right), \tag{2}$$

где C_k^t – расстояние между k-ми характеристиками элементов, w_k^t – вес для k-ой характеристики. Функция стоимости связи имеет следующий вид:

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} \left(w_k^c C_k^c \left(u_{i-1}, u_i \right) \right), \tag{3}$$

где C_k^c – расстояние между k-ми характеристиками элементов, w_k^c – вес для k-ой характеристики.

При решении задачи поиска оптимальной последовательности звуковых элементов ключевым является выбор признаков, на основе которых вычисляются функции стоимости. В процессе реализации гибридной системы синтеза речи был проведен отбор таких характеристик на основе оценки качества звучания результирующего сигнала. В итоге были выработаны следующие наборы признаков:

- для стоимости замены: длительность аллофона, энергия аллофона, частота основного тона (4 точки для ударных, две – для безударных), тип контекстов (соответствие соседних аллофонов элемента из базы его окружению в синтезируемом предложении), тип интонационной модели (повествовательная, вопросительная, восклицательная), позиция аллофона в слоге (слове, предложении), тип слога (ударный, безударный);
- для стоимости связи: значение частоты основного тона и его производной на границах аллофона, значение энергии и ее производной на границах аллофона, значение мел-частотных кепстральных коэффициентов (MFCC) на границах аллофона, непрерывность фрагмента (искусственное уменьшение веса подряд идущих аллофонов, обеспечивающих синтез требуемого фрагмента текста).

В третьей главе приводится реализация системы синтеза речи, основанной на гибридном методе, объединяющем в себе два независимых подхода: синтез речи на основе метода Unit Selection и синтез речи на основе скрытых марковских моделей. В этом случае сначала на основе скрытых марковских моделей выполняется формирование модели голоса диктора, которая в процессе синтеза речи позволяет определить акустические характеристики (частоту основного тона и длительность) аллофонов, согласно которым происходит выбор множества элементов из звуковой базы данных. Далее методом Unit Selection определяется их оптимальная последовательность на основе стоимостной функции (1).

Идеологически и структурно систему можно разделить на 2 части (рисунок 2): подготовка звуковой базы данных и синтез речи.

В основе подготовки звуковой базы данных лежит речевой корпус, состоящий из набора речевых файлов, каждый из которых содержит произнесение одного предложения, и соответствующего ему набора файлов разметки,

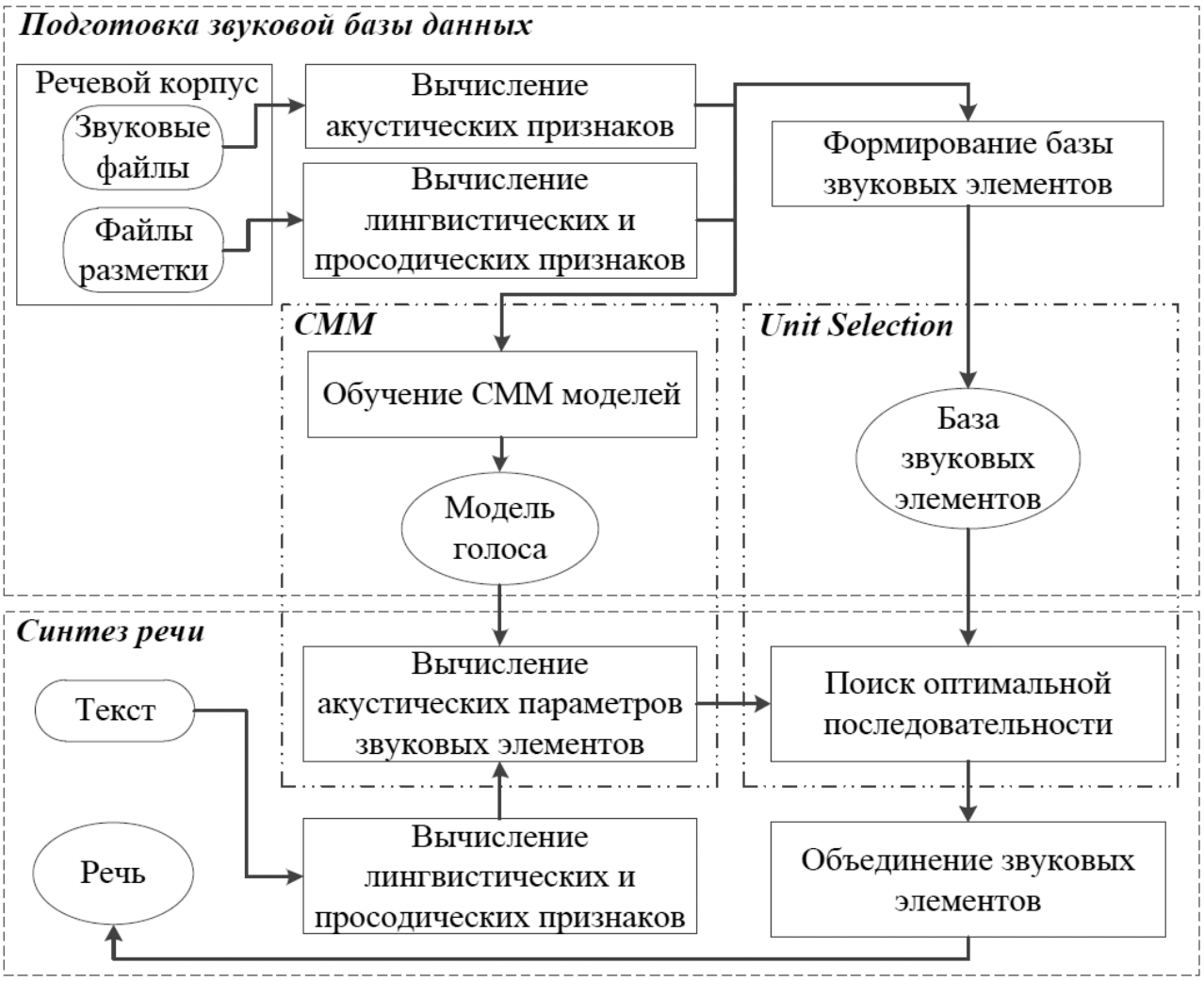

Рис. 2: Общая схема гибридной системы синтеза речи

каждый из которых содержит подробную информацию о предложении в речевом файле.

Процедура моделирования параметров голоса начинается с генерации набора характеристик для всех речевых файлов. Каждый такой набор описывает короткий участок сигнала (фрейм) длиной 25 мс. В качестве характеристик автором выбраны следующие параметры.

- Последовательность векторов MFCC коэффициентов, рассчитанных для каждого фрейма, где каждый вектор $\{c_1, ..., c_K\}$ характеризует спектральную огибающую сигнала; K – общее количество фреймов; количество коэффициентов в каждом векторе равно 25.
- Последовательность значений частоты основного тона $\{F0_1, ..., F0_K\}$, включая информацию о том, является ли участок вокализованным или нет.

На следующем шаге происходит вычисление лингвистических и просодических признаков, представленных в предыдущем разделе, для каждого аллофона на основе файлов разметки. Далее для каждого аллофона происхо-

дит создание прототипов СММ-моделей. Каждая модель имеет N состояний, где переход из каждого возможен только само в себя или в следующее состояние. В разработанном автором методе $N = 5$. Каждый выходной вектор наблюдений $\mathbf{o}^i$ содержит 4 потока $\mathbf{o}^i = [\mathbf{o}_1^{i\,T}, \mathbf{o}_2^{i\,T}, \mathbf{o}_3^{i\,T}, \mathbf{o}_4^{i\,T}]^T$, где первый поток содержит значения коэффициентов MFCC, их первых и вторых производных, второй поток – значение частоты основного тона (ЧОТ), третий поток – значение первой производной ЧОТ, четвертый поток – значение второй производной ЧОТ.

Вектор наблюдения $\mathbf{o}^i$ является выходом состояния n СММ-модели, вероятность которого определяется следующим выражением:

$$\beta_n\left(\mathbf{o}^i\right) = \prod_{j=1}^{4}\left(\sum_{l=1}^{R_j}\omega_{njl}N(\mathbf{o}_j^i;\boldsymbol{\mu}_{njl},\boldsymbol{\Sigma}_{njl})\right), \tag{4}$$

где $N(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ определяет гауссово распределение с вектором средних $\boldsymbol{\mu}$ и матрицей ковариации $\boldsymbol{\Sigma}$, ω_{njl} – вес l-ой компоненты смеси j-го потока выходного вектора состояния n с количеством компонентов в смеси, равным R_j.

Для каждой k-ой СММ-модели длительность N состояний соответствует вектору $\mathbf{d}^k = [\mathbf{d}_1^k, ..., \mathbf{d}_N^k]^T$, где $\mathbf{d}_n^k$ определяет длительность n-го состояния. Каждый вектор длительностей моделируется N-размерным однокомпонентным гауссовым распределением.

Затем выходные вероятности моделей спектральных параметров (MFCC и ЧОТ) и длительностей переоцениваются на основе алгоритма Баума-Велша.

В заключении процесса построения модели голоса выполняется кластеризация состояний СММ-моделей для спектральных характеристик и параметров длительностей на основе деревьев решений. Выбор оптимального разбиения осуществляется на основе максимизации функции $L(U)$ (5), вычисляемой для каждого из подразбиений.

$$L(U) = -\frac{1}{2}T\left(L + L\log 2\pi + \log|\bar{\boldsymbol{\Sigma}}|\right), \tag{5}$$

где U – множество состояний $\{S_1, S_2, ..., S_M\}$, $S_i = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $T = \sum_{i=1}^{M} O_i$, O_i – количество повторений состояния S_i, L – размерность вектора признаков,

$$\boldsymbol{\mu}_i = \{\mu_{i1}, ..., \mu_{iL}\}, \boldsymbol{\Sigma}_i = \begin{Bmatrix} \sigma_{i1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{iL}^2 \end{Bmatrix}, \bar{\boldsymbol{\Sigma}}_i = \begin{Bmatrix} \bar{\sigma}_{i1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \bar{\sigma}_{iL}^2 \end{Bmatrix},$$

$\bar{\sigma}_i^2 = (\hat{\sigma}_i^2 - \hat{\mu}_i/T)/T$, $\hat{\mu}_i = \sum_{j=1}^{M}\mu_{ij}O_j$, $\hat{\sigma}_i^2 = \sum_{j=1}^{M}(\sigma_{ij}^2 + \mu_{ij}^2)O_j$.

Выполнение данного шага дает возможность вычислять параметры тех звуковых единиц, которые отсутствуют в обучающем речевом корпусе, что позволяет получать синтезированную речь даже при наличии небольшого количества речевого материала. В результате полученная модель голоса представляет собой $N+1$ различных деревьев: N – для хранения значений частоты основного тона, их первой и второй производной и одно – для хранения значений длительностей. Звуковые элементы упаковываются в единую базу данных, обеспечивающую поиск по целевым характеристикам, таким как имя аллофона, имена аллофонов слева и справа, коэффициенты MFCC на границах, энергия на границах, частота основного тона на границах и длительность аллофона.

Процедура создания модели голоса (рисунок 2), состоящая из процедуры извлечения признаков, оценки параметров CMM модели каждого типа звукового элемента, переоценки данных параметров и кластеризации состояний, является достаточно трудоемкой и требует больших временных затрат (рисунок 7). Для оптимизации данного процесса были разработаны параллельные алгоритмы, позволяющие существенно повысить эффективность (рисунок 3): конвейеризация обработки файлов на этапе извлечения признаков, параллельная независимая оценка параметров каждого типа звукового элемента, параллельная независимая кластеризация каждого потока состояний CMM моделей, конвейеризация оценки оптимального разбиения.

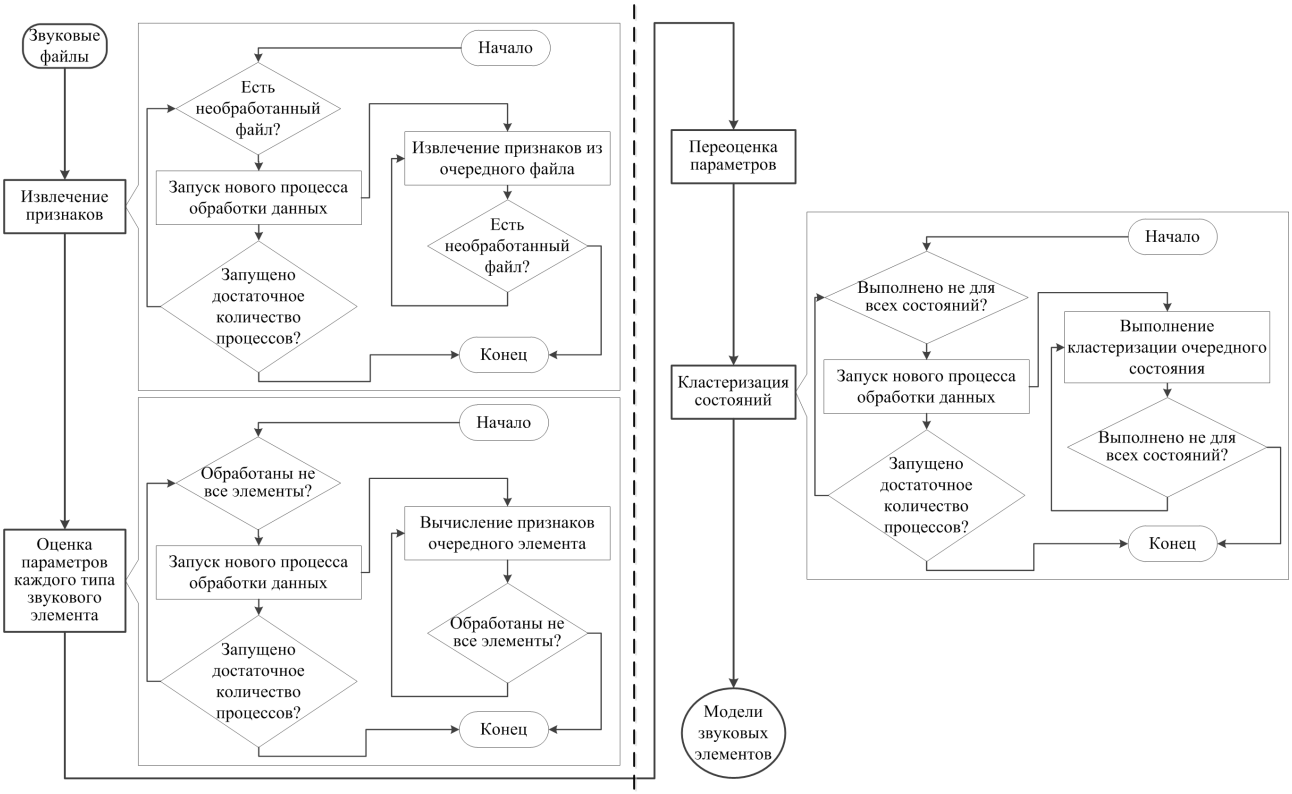

Рис. 3: Схема распараллеливания процедуры создания модели голоса

Синтез речи выполняется на основе текста, подаваемого на вход системы без какой-либо предварительной ручной обработки. На основе текстовой информации происходит определение аллофонной последовательности для синтеза и вычисление лингвистических и просодических признаков для каждого аллофона, характеризующих их положение в данном предложении. Тип и структура признаков аналогична тем, что вычисляются на этапе подготовки звуковой базы данных. На основе лингвистической и просодической информации производится определение физических признаков каждого аллофона на основе модели голоса. В качестве акустических метрик выступают следующие параметры: частота основного тона (несколько точек на каждый вокализованный аллофон), значение энергии и значение длительности.

Далее происходит выбор групп наиболее подходящих звуковых элементов из базы на основе значений акустических характеристик, вычисленных для каждого аллофона; формирование аллофонной решетки, описывающей синтезируемое предложение, и поиск на ней оптимального пути, т.е. формирование последовательности звуковых элементов. На заключительном шаге происходит объединение выбранной последовательности элементов в единый звуковой поток, на выходе представляющий собой синтезированную речь. В отличие от общепринятой практики стыковки звуковых элементов с перекрытием на границах, в данной работе для повышения естественности автором применяется анализ объединяемых элементов и типов контекстов, в которых они были подобраны методом Unit Selection. Так, стыковка вокализованных элементов с невокализованными осуществляется в области невокализованного элемента, а стыковка элемента из близкого контекста с элементом из дальнего - в области элемента из дальнего контекста.

Заключительным этапом является модификация речи по частоте основного тона и длительности для обеспечения требуемых темпо-ритмических характеристик выбранных элементов. В основе разработанного алгоритма лежит модель линейного предсказания. Применение данной модели позволяет разделить речь на характеристики вокального тракта и характеристики голосовых связок. Модификация по частоте основного тона выполняется путем изменения расстояния между «пиками» в сигнале возбуждения, а длительности - путем повторения или редукции отдельных его участков. Данный подход позволил существенно снизить уровень искажений, вносимых при модификации в речевой сигнал (по результатам экспертной оценки), и расширить область применения алгоритма: пределы модификации ЧОТ от 0,5 до 2 раз, пределы модификации длительности от 0,5 до 4 раз от исходных значений.

Программные средства синтеза естественной русской речи (рисунок 4) состоят из парсера входных текстовых потоков различных форматов и четырех основных процессоров верхнего уровня (лингвистический процессор, про-

содический процессор, фонетический процессор, акустический процессор). Каждый процессор имеет свою специфическую базу данных и наборы правил выделения информации, необходимой для синтеза, из входного потока. Такое разделение позволяет обеспечить максимальную независимость процессоров и легкую их замену. Все процессоры предоставляют одинаковый программный интерфейс, обмен данными выполняется в XML-формате.

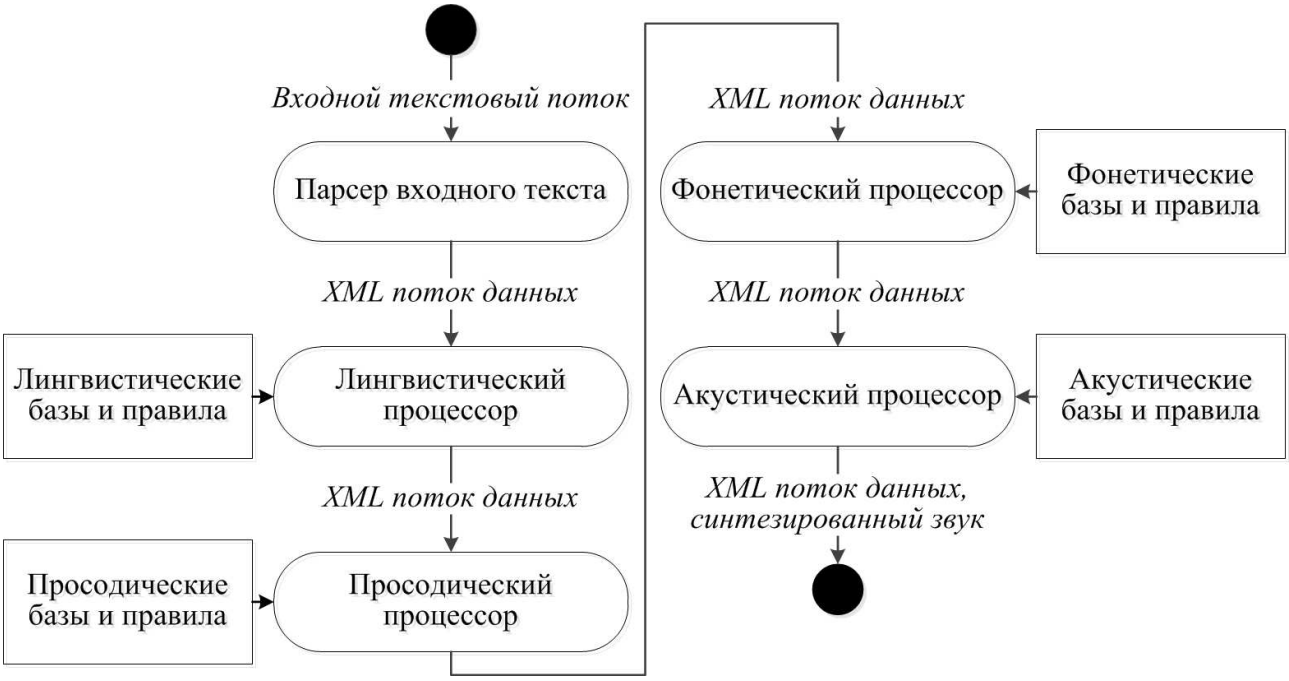

Рис. 4: Общая схема системы синтеза русской речи

Каждый из разработанных процессоров инкапсулирует в себе реализацию следующей функциональности: лингвистический – нормализация текста, расшифровка числительных, аббревиатур, транслитерация слов, написанных латиницей; просодический – определение границ предложений, вычисление признаков текстовых элементов; фонетический – моделирование физических параметров звуковых элементов; акустический – поиск оптимальной последовательности звуковых элементов методом Unit Selection, их модификация и объединение в единый звуковой поток, представляющий собой синтезированную речь. В ходе выполнения диссертационной работы автором были разработаны следующие компоненты: модуль разрешения неоднозначностей произношения слов (подпроцессор лингвистического процессора), модуль вычисления признаков текстовых элементов (подпроцессор просодического процессора), фонетический и акустический процессоры.

В четвертой главе представлены результаты экспериментальных исследований и технические характеристики реализованной системы. На рисунках 5, 6 и 7 приведены, соответственно, осциллограммы, спектрограммы и графики динамики частоты основного тона для фразы «это очень важно!».

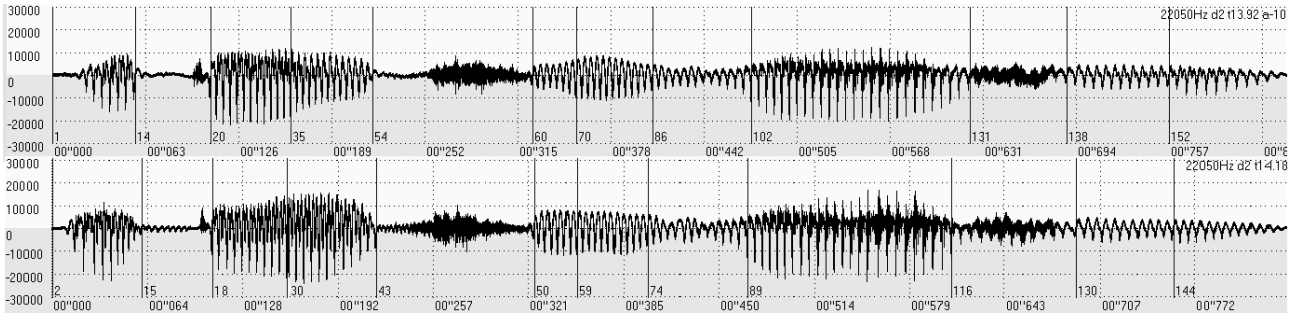

Рис. 5: Осциллограммы фразы «это очень важно!», записанной реальным диктором (сверху) и синтезированной системой (снизу)

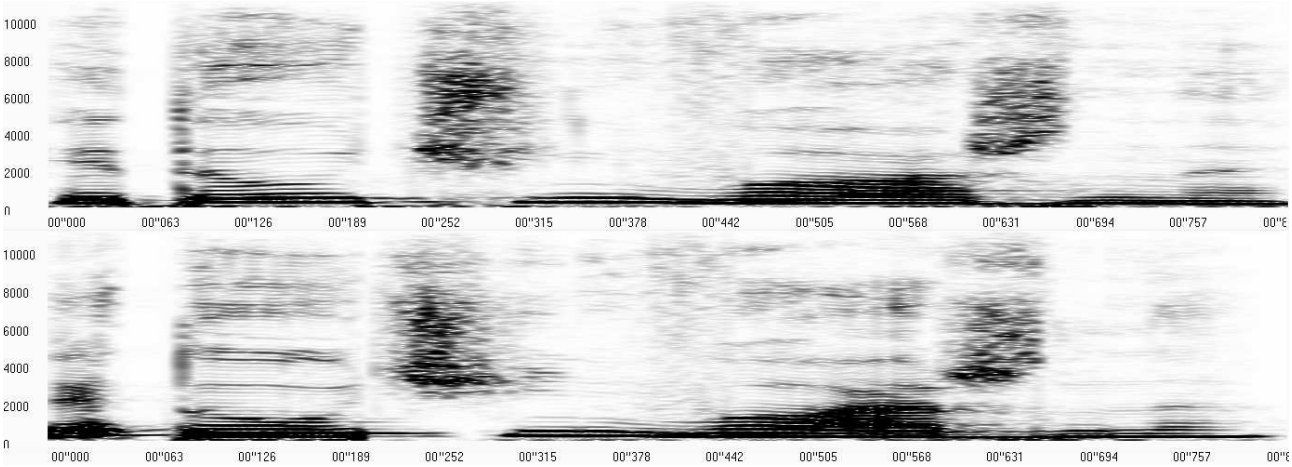

Рис. 6: Спектрограммы фразы «это очень важно!», записанной реальным диктором (сверху) и синтезированной системой (снизу)

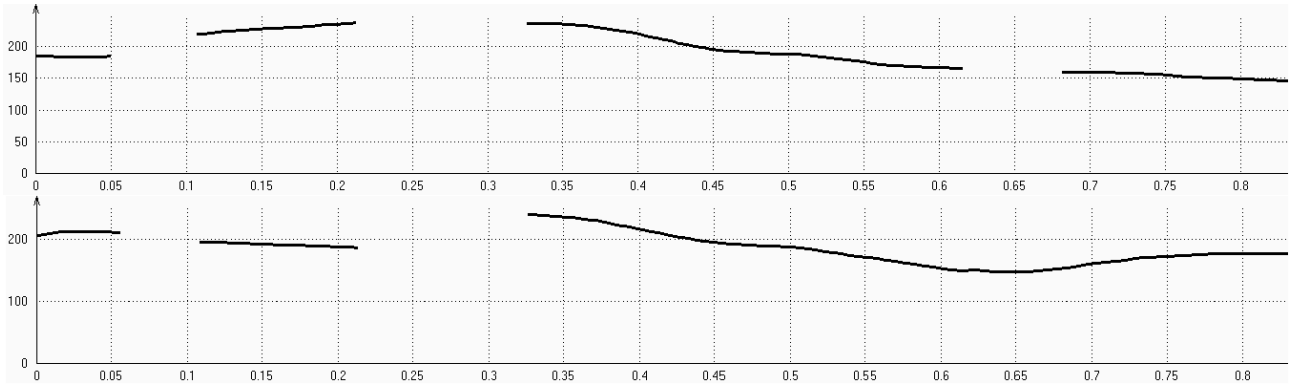

Рис. 7: Графики динамики ЧОТ фразы «это очень важно!», записанной реальным диктором (сверху) и синтезированной системой (снизу)

На приведенных рисунках в верхней части представлены данные для естественной фразы, записанной реальным диктором, а в нижней – ее синтезированный вариант. Следует отметить, что синтезируемая фраза не была включена в обучающую выборку.

На основе данных диаграмм можно сделать вывод, что синтезированная фраза имеет такие же темпо-ритмические и спектральные характеристики, что и ее эквивалент, произнесенный диктором. Это достигается за счет моделирования значений этих характеристик на основе скрытых марковских моделей.

В таблице 1 представлены результаты сравнения показателей естественности речи (значения в интервале от 0 до 5, где 5 определяет максимальную оценку естественности) для систем синтеза на английском языке, ставших лучшими по итогам соревнований Blizzard Challenge 2010, с системой синтеза речи на русском языке, представленной в данной работе, и системой на основе метода Unit Selection, лежащей в основе гибридного подхода. Данные в таблице усреднены по всем дикторам. Как видно из результатов эксперимента, применение гибридного подхода, как и в случае с английским языком, позволило улучшить показатели естественности синтезированной речи.

Таблица 1: Показатели естественности систем синтеза

Язык	Тип подхода к синтезу		Естественная речь
	Unit Selection	Гибридный подход	
Английский	3.8	4.2	4.8
Русский	4.0	4.3	4.8

Производительность системы определяется двумя факторами: 1) время обучения моделей; 2) скорость синтеза фразы. Оценка производилась на персональном компьютере следующей конфигурации:

- параметры процессора: Intel Xeon CPU X5365 3.00 ГГц;
- объем оперативной памяти: 16 Гб;
- операционная система: MS Windows Server 2008 Standard x64.

Результаты замеров по первому критерию приведены на рисунке 8. На данном графике представлена зависимость времени обучения моделей от объема звуковой базы данных, где система A – инструмент с открытыми исходными кодами HTS (http://hts.sp.nitech.ac.jp), предназначенный для обучения скрытых марковских моделей; B – система, представленная в данной работе. Как видно из результатов оценки, время, требуемое для обучения моделей по базе объемом 8 часов, составляет порядка 24 часов, что практически на порядок быстрее аналогичной системы HTS.

Второй критерий характеризуется зависимостью производительности системы (в единицах реального времени (RT)) от объема речевой базы данных. Наиболее трудоемкой операцией на данном этапе является поиск опти-

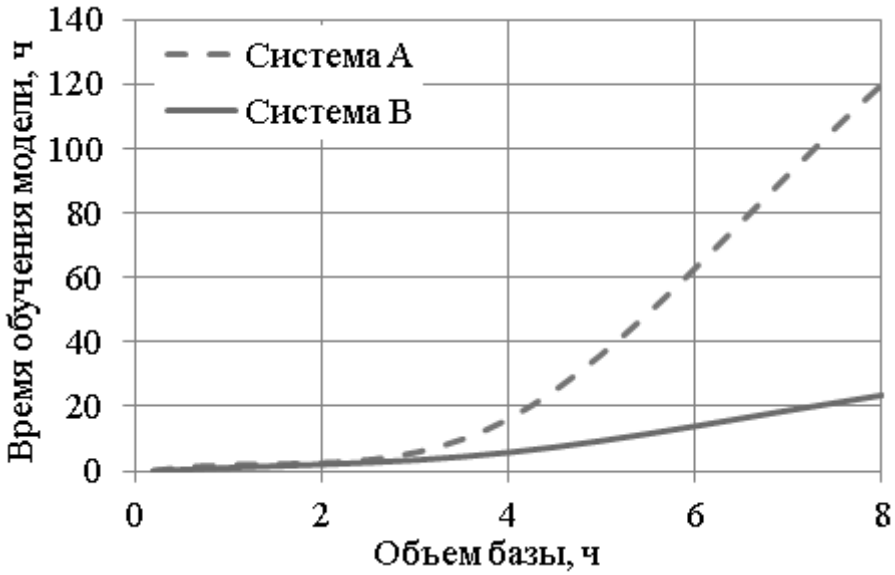

Рис. 8: Зависимость времени обучения моделей от объема речевой базы

мального пути на графе при выборе элементов, что с ростом их количества ведет к снижению скорости работы. Стоит отметить, что даже при объеме звуковой базы данных, равном десяти часам, производительность системы составляет 201 RT, что полностью соответствует требованиям реальных приложений к подобным системам и сравнимо с показателями систем синтеза для других языков.

Заключение

В ходе проведенных исследований была разработана гибридная система синтеза русской речи по тексту, в основе которой лежат скрытые марковские модели и метод Unit Selection. Результаты испытаний показали, что по показателям естественности звучания данная система является лучшей среди систем синтеза на русском языке, при этом полностью удовлетворяя диктуемым реальными приложениями требованиям по производительности (скорости работы и занимаемому объему памяти). Разработанная система успешно себя зарекомендовала в различных научно-исследовательских и опытно-конструкторских разработках, а также коммерческих решениях компании ООО «ЦРТ» как в качестве самостоятельного продукта, так и в составе других, например, системы голосового самообслуживания.

В диссертации получены следующие результаты.

1. Создана методика подготовки речевого корпуса, включающая формирование текстового корпуса, автоматический контроль параметров записи фонограмм, автоматическую разметку звукового материала.
2. Выбран набор признаков звуковой единицы русского языка и набор критериев поиска последовательности звуковых элементов методом Unit Selection.

3. Разработана методика создания модели голоса.
4. Реализовано масштабируемое ПО обучения моделей голоса.
5. Разработаны алгоритмы и реализовано ПО модификации частоты основного тона, энергии и длительности и стыковки звуковых элементов.
6. Разработаны программные средства синтеза русской речи, основанной на совместном использовании скрытых марковских моделей и метода Unit Selection.

Статьи в журналах из перечня ВАК

1. Чистиков П.Г., Рыбин С.В. Проблемы естественности речевого сигнала в системах синтеза // Журнал «Компьютерные инструменты в образовании». – 2011. – Вып. 1. – С. 22-30.
2. Чистиков П.Г., Хомицевич О.Г. Автоматическое определение границ предложений в потоковом режиме в системе распознавания русской речи // Вестник МГТУ им. Н.Э. Баумана Сер. Приборостроение. – 2011. – Вып. S. – С. 117-125.
3. Чистиков П.Г. Технология синтеза русской речи на основе скрытых марковских моделей // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – Вып. 3. – С. 151-152.
4. Чистиков П.Г., Корольков Е.А., Таланов А.О., Соломенник А.И. Гибридная технология синтеза русской речи на основе скрытых марковских моделей и алгоритма Unit Selection // Журнал «Известия вузов. Приборостроение». – 2013. – Вып. 2. – С. 33-38.
5. Соломенник А.И., Чистиков П.Г., Рыбин С.В., Таланов А.О., Томашенко Н.А. Автоматизация процедуры подготовки нового голоса для системы синтеза русской речи // Журнал «Известия вузов. Приборостроение». – 2013. – Вып. 2. – С. 29-32.
6. Соломенник А.И., Таланов А.О., Чистиков П.Г., Соломенник М.В., Хомицевич О.Г. Проблемы и решения задачи оценки качества синтезированной речи // Журнал «Известия вузов. Приборостроение». – 2013. – Вып. 2. – С. 38-42.

Другие публикации

7. Главатских И.А., Чистиков П.Г., Таланов А.О. Метод модификации физических параметров речевого сигнала на основе периодосинхронного Фурье-анализа // Труды XXXVIII международной филологической конференции. – 2008. – С. 47-62.
8. Аничкин И.М., Чистиков П.Г. Формализация правил автоматического снятия омонимии в системе синтеза речи по тексту // Труды XXXVIII международной филологической конференции. – 2008. – С. 29-45.

9. Продан А.И., Чистиков П.Г., Таланов А.О. Система подготовки нового голоса для системы синтеза «VITALVOICE» // Сборник трудов международной конференции по компьютерной лингвистике «Компьютерная лингвистика и интеллектуальные технологии». – 2010. – Вып. 9(16). – С. 394-399.
10. Смирнова Н.С., Чистиков П.Г. Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задачах в области речевых технологий // Сборник трудов международной конференции по компьютерной лингвистике «Компьютерная лингвистика и интеллектуальные технологии». – 2011. – Вып. 10(17). – С. 632-643.
11. Chistikov P.G. Pitch-scale modification in text-to-speech systems // Proceedings of the IEEE North West Russia Section. – 2011. – P. 37-42.
12. Chistikov P., Khomitsevich O. On-line automatic sentence boundary detection in a Russian ASR system // SPECOM 2011 International Conference. – 2011. – P. 112-117.
13. Chistikov P., Talanov A. High Quality Pitch-Scale Modification in Speech Generation Systems // SPECOM 2011 International Conference. – 2011. – P. 367-372.
14. Smirnova N., Chistikov P. Statistics of Russian Monophones and Diphones // SPECOM 2011 International Conference. – 2011. – P. 218-223.
15. Чистиков П.Г. Моделирование параметров русской речи в системе синтеза // Сборник тезисов докладов конгресса молодых ученых, Выпуск 2. Труды молодых ученых / Главный редактор д.т.н., проф. В.О. Никифоров. – СПб: НИУ ИТМО, 2012. – С. 227-228.
16. Chistikov P.G., Korolkov E.A. Data-driven Speech Parameter Generation For Russian Text-to-Speech System // Proceedings of the International Conference on Computational Linguistics «Computational Linguistics and Intellectual Technologies». – 2012. – № 11(18). – P. 103-111.
17. Solomennik A.I., Chistikov P.G. Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system // Proceedings of the International Conference on Computational Linguistics «Computational Linguistics and Intellectual Technologies». – 2012. – № 11(18). – P. 607-615.

Тиражирование и брошюровка выполнены
в центре «Университетские телекоммуникации»
Санкт-Петербург, Саблинская ул., 14, тел. +7(812)233-46-69.
Объем 1,0 у.п.л. Тираж 100 экз.