|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_«Информатика и системы управления»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_«Программное обеспечение ЭВМ и информационные технологии»\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

**Разработка алгоритма для синтеза речи по заранее неизвестному тексту на русском языке**

Студент \_ИУ7-75Б\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_**А. П. Овчинникова**\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_**Т. Н. Романова**\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2020 г.*

**Оглавление**

[Введение 3](#_Toc63883934)

[1. Аналитический раздел 5](#_Toc63883935)

[1.1. Задача преобразования текста в речь 5](#_Toc63883936)

[1.2. Обзор существующих систем преобразования текста в речь 5](#_Toc63883937)

[1.2.1. Amazon Polly 5](#_Toc63883938)

[1.2.2. IBM Watson Text to Speech 6](#_Toc63883939)

[1.2.3. Azure Text to Speech API 6](#_Toc63883940)

[1.2.4. Google Cloud Text-to-Speech 7](#_Toc63883941)

[1.2.5. Обоснование необходимости разработки нового продукта 8](#_Toc63883942)

[1.3. Методы синтеза речи 8](#_Toc63883943)

[1.3.1. Параметрический синтез 8](#_Toc63883944)

[1.3.2. Компилятивный синтез 9](#_Toc63883945)

[1.3.3. Синтез речи по фонетическим правилам 9](#_Toc63883946)

[1.3.4. Выбор метода синтеза речи 10](#_Toc63883947)

[1.4. Выбор исходных элементов при создании базы данных для синтеза речи по правилам 10](#_Toc63883948)

[1.5. Мини- и макси-наборы аллофонов для синтеза русской речи 11](#_Toc63883949)

[2. Конструкторский раздел 13](#_Toc63883950)

[2.1. Требования к программе 13](#_Toc63883951)

[2.2. Описание входных и выходных данных и ограничений 13](#_Toc63883952)

[Список литературы 16](#_Toc63883953)

Введение

Речевое общение является доминирующим способом человеческих социальных связей и обмена информацией. Эта тенденция находит отражение и во взаимодействии человек-машина. В современном мире появилась необходимость создания систем разговорного языка. Такие системы сочетают в себе как распознавание речи, так и ее синтез. Создание подобных жизнеспособных систем уже давно привлекает внимание ученых и инженеров во всем мире. Успешное развитие этой технологии повысит доступность компьютеров и автоматизированных систем для широкого круга пользователей.

Под автоматическим синтезом речи понимается технология, позволяющая преобразовать текстовую информацию в звучащую речь. Первые синтезаторы речи, механические, появились еще в 18 веке. Электрические синтезаторы стали создаваться в 19 веке после того, как была разработана резонаторная теории Гельмгольца. С тех пор в технологии синтеза речи прошло три поколения, и современные синтезаторы третьего поколения позволяют получать синтезированную речь, наиболее приближенную к естественной.

В настоящее время технология автоматического синтеза находит широкое применение в таких отраслях, как телекоммуникации, мобильные устройства, автомобильная индустрия, компьютеризованные системы, образовательные системы и многих других. Подавляющее большинство систем разговорного языка создаются как универсальные, поддерживающие широкий набор языков. Однако стремление к универсальности может нанести серьезный ущерб качеству синтезируемой речи, потому как каждая языковая система обладает уникальными чертами, которые стираются при попытке создать универсальную систему для всех (или многих) языков. Для создания качественного синтезатора речи необходимо, помимо знаний в программировании и в обработке звука на компьютере, необходимо также обратиться к таким областям лингвистики, как фонетика и просодика.

Цель данной работы – разработка алгоритма для синтеза речи по заранее неизвестному тексту на русском языке – призвана решить обозначенную проблему.

Для достижения поставленной цели необходимо решить следующие задачи:

* провести анализ существующих методов синтеза речи;
* выбрать наиболее подходящий метод синтеза речи;
* изучить фонетические особенности русского языка;
* собрать базу данных, содержащую необходимый набор исходных элементов для синтеза;
* спроектировать программное обеспечение, реализующее выбранный метод синтеза речи;
* выбрать технические средства для реализации спроектированного программного обеспечения;
* реализовать программное обеспечение;
* провести исследование для оценки качества речи, синтезируемой с помощью разработанного программного обеспечения.

1. Аналитический раздел

В данном разделе рассматриваются существующие системы преобразования текста в речь (Text To Tpeech Systems), проводится обзор методов синтеза речи, а также дается краткое описание фонетических особенностей русского языка.

* 1. Задача преобразования текста в речь

Программы синтеза речи преобразуют входные текстовые сообщения в устный вывод, автоматически генерируя синтетическую речь. Синтез речи часто называют преобразованием текста в речь (TTS). Под системами автоматического синтеза речи понимают системы, преобразующие письменный орфографический текст и другую информацию в звучащую речь [1].

1.2. Обзор существующих систем преобразования текста в речь

Многие крупные компании, такие как Mirosoft, Amazon и другие активно занимаются разработкой собственных text-to-speech систем, позволяющих преобразовать в речь неизвестные заранее входные данные. Каждая такая система обладает своими достоинствами и недостатками.

1.2.1. Amazon Polly

Amazon Polly - это облачный сервис Amazon Web Services, дочерней компании Amazon.com, который преобразует текст в речь, использующий технологии глубокого обучения [1]. Сервис позволяет работать с различными языками и включает в себя несколько голосов для каждого языка. Основные достоинства и недостатки Amazon Polly приведены в Таблица 1.

Таблица 1. Сервис Amazon Polly.

|  |  |
| --- | --- |
| Достоинства | Недостатки |
| * Поддержка различных языков, в том числе русского. * Поддержка одного или более голосов для одного языка. * Предоставляет возможность управления такими аспектами речи, как произношение, громкость, высота голоса, скорость и т. д. * Подробная документация. * Наличие обучающих пособий от разработчика. * Бесплатная ограниченная версия. | * Является платным сервисом, стоимость использования зависит от количества символов в тексте. * Нельзя заранее проверить качество синтезированной речи на необходимом тексте. Можно только прослушать заранее заготовленные примеры на сайте. * Сервис имеет закрытый исходный код. * Необходимо обучение для использования продукта. * Обучающие пособия и документация на английском языке. |

1.2.2. IBM Watson Text to Speech

IBM Watson Text to Speech – облачный сервис, который позволяет преобразовывать письменный текст в естественную речь на разных языках и разных голосах [2]. Сервис основан на технологии нейронного голоса. Основные достоинства и недостатки IBM Watson Text to Speech приведены в Таблица 2.

Таблица 2. Сервис IBM Watson Text to Speech.

|  |  |
| --- | --- |
| Достоинства | Недостатки |
| * Поддержка различных языков. * Необходимо обучение для использования продукта. * Обучающие пособия и документация на английском языке. * Поддержка различных голосов. Помимо стандартных голосов есть возможность создать уникальный пользовательский голос. * Бесплатная ограниченная версия. | * Нет поддержки русского языка. * Является платным сервисом, стоимость использования зависит от количества символов в тексте. * Нельзя заранее проверить качество синтезированной речи на необходимом тексте. Можно только прослушать заранее заготовленные примеры на сайте. * Сервис имеет закрытый исходный код. * Необходимо обучение для использования продукта. * Обучающие пособия и документация на английском языке. |

1.2.3. Azure Text to Speech API

Azure Text to Speech API – служба, позволяющая приложениям, средствам или устройствам преобразовывать текст с помощью языка разметки синтеза речи (SSML) [3]. Можно использовать речевой пакет SDK или REST API. Основные достоинства и недостатки Azure Text to Speech API приведены в Таблица 3.

Таблица 3. Служба Azure Text to Speech API.

|  |  |
| --- | --- |
| Достоинства | Недостатки |
| * Поддержка различных языков, в том числе русского. * Поддержка различных голосов. Помимо стандартных голосов есть возможность создать уникальный пользовательский голос. * Подробная документация на русском языке. * Бесплатная пробная подписка. * Можно заранее проверить качество синтезированной речи на необходимом тексте. | * Является платной службой, стоимость использования зависит от количества символов в тексте. * Имеет закрытый исходный код. * Необходимо обучение для использования продукта. |

1.2.4. Google Cloud Text-to-Speech

Google Cloud Text-to-Speech позволяет разработчикам синтезировать естественную речь с 30 голосами, доступными на нескольких языках и в разных вариантах [4]. Google Cloud Text-to-Speech основан на исследованиях DeepMind. Сервис позволяет преобразовывать текст с помощью языка разметки синтеза речи (SSML). Можно использовать REST API. Основные достоинства и недостатки Google Cloud Text-to-Speech представлены в Таблица 4.

Таблица 4. Google Cloud Text-to-Speech.

|  |  |
| --- | --- |
| Достоинства | Недостатки |
| * Можно заранее проверить качество синтезированной речи на необходимом тексте. * Подробная документация. * Можно заранее проверить качество синтезированной речи на необходимом тексте. * Поддержка различных голосов. Помимо стандартных голосов есть возможность создать уникальный пользовательский голос. Для стандартных голосов можно изменять тон, громкость и скорость. * Можно выбрать формат выходного аудиофайла. * Оптимизация под тип динамика. * Бесплатный пробный период. | * Документация на английском языке. * Является платным сервисом, стоимость использования зависит от количества символов в тексте. * Имеет закрытый исходный код. * Необходимо обучение для использования продукта. |

1.2.5. Обоснование необходимости разработки нового продукта

Нами были рассмотрены лишь некоторые из наиболее известных text-to-speech систем. На сегодняшний день можно найти большое количество подобных систем, но все они в основном обладают схожими недостатками. Проанализировав существующие на рынке решения для преобразования текста в речь, можно сделать вывод, что разработка нового продукта имеет смысл.

Таким образом, разрабатываемый программный продукт должен сочетать в себе следующие особенности:

* поддержка русского языка;
* учет фонетических особенностей русского языка;
* открытый исходный код;
* бесплатное распространение;
* простота использования, документация на русском языке.

Главной чертой программы является поддержка русского языка для ее распространения в русскоязычном сообществе.

1.3. Методы синтеза речи

Существует несколько алгоритмов генерации речи. Выбор алгоритма зависит, в первую очередь, от задачи, для которой он будет использоваться.

В настоящее время выделяют три группы методов синтеза речи:

* параметрический синтез;
* компилятивный синтез;
* синтез речи по фонетическим правилам.

1.3.1. Параметрический синтез

Параметрический синтез применяется в вокодерных системах. Вокодеры синтезируют речь на основе сигнала, представленного набором непрерывно изменяющихся во времени параметров. Эти параметры и управляют синтезатором речи.

Параметрический метод используется, когда набор текстовых сообщений, которые необходимо озвучить, ограничен и заранее известен. В этом случае качество синтезируемой речи может быть очень высоким. Недостатком метода является невозможность его применения к произвольному набору заранее неизвестных текстовых сообщений.

1.3.2. Компилятивный синтез

Суть данного метода заключается в компиляции речи из предварительно заготовленных и записанных исходных элементов синтеза – слов, фраз, предложений. Минимальным размером элементов при этом является слово. Таким образом, возможности компилятивного метода ограничены объемом словаря.

Главная проблема метода – необходимость в большом объеме памяти для хранения предварительно записанного словаря, при этом содержание и вариативность синтезируемой речи ограничена этим словарем. Несмотря на это компилятивный синтез получил большое распространение и широкое практическое применение в узконаправленных системах с ограниченным набором озвучиваемых сообщений. Например, объявления в метро: «Следующая станция …», «Поезд следует до станции …». Названия станций записаны заранее и просто соединяются с объявлением. Другим примером являются справочные службы сотовых операторов при озвучивании информации об остатке средств на счету абонента.

Этот метод является самым простым и качество выходной речи зависит только от способа записи минимальных элементов синтеза.

1.3.3. Синтез речи по фонетическим правилам

Более сложными, но худшими по качеству являются алгоритмы, разделяющие речь на более мелкие части. Метод синтеза речи по фонетическим правилам, основываясь на размере минимальных исходных элементов синтеза, делят на следующие виды:

* микросегментный;
* аллофонный;
* дифонный;
* полуслоговый;
* слоговый;
* синтез из различных единиц разного размера.

Синтез речи по фонетическим правилам позволяет синтезировать речь по заранее неизвестному тексту. Однако синтезированная таким образом речь может отличаться от естественной, потому что при склейке элементов на их границах возникают искажения. Чем длиннее становятся минимальные блоки, тем больше таких элементов в словаре и качество синтезируемой речи увеличивается вместе с требуемой памятью.

Кроме того, при таком подходе сложно управлять просодическими характеристиками синтезированной речи, поскольку интонационные и акустические характеристики слова зависят от его места в предложении.

1.3.4. Выбор метода синтеза речи

В качестве способа генерации речи был выбран синтез по фонетическим правилам, потому как это единственный метод, позволяющий озвучить произвольный заранее неизвестный текст. Для реализации данного подхода необходимо создать фонетико-акустическую базу данных, содержащую большое количество минимальных исходных элементов синтеза.

1.4. Выбор исходных элементов при создании базы данных для синтеза речи по правилам

Выбор исходных единиц синтеза существенно влияет на качество синтезируемой речи, отсюда возникает задача оптимального выбора единиц, с которыми будет работать синтезатор речи.

Необходимо учитывать, что фонетические единицы могут изменяться под действием языковых правил (ассимиляция, диссимиляция, оглушение и изменение гласных звукотипов в зависимости от ударения), а также под действием моторной программы (аккомодация). Аккомодацией (коартикуляцией) называется одновременное осуществление артикуляционных движений, относящихся к разным артикуляционным жестам, что облегчает артикуляционный переход от одного звука к другому [7]. При произнесении последовательности звуков возникают переходные участки, так как для произнесения следующего звука необходимо перестроить речевой аппарат. Переходные участки могут быть короткими по длительности, однако несут в себе важную для восприятия речи информацию. Например, информация о твердости или мягкости согласного часто содержится в следующем за ним гласным. Твердость/мягкость согласных в русском языке является смыслоразличительной характеристикой (*гроза – грозя*). Очевидно, что для качественного синтеза речи необходимо учитывать переходные участки. Для этого необходимо в качестве исходных элементов выбирать более крупные чем отдельные звуки элементы. Такими элементами являются аллофоны.

Кроме аллофонного, существует дифонный синтез речи. Дифоном называют сегмент речи между серединами соседних аллофонов. Однако при дифонном синтезе иногда бывает трудно выделить границу деления на дифоны, что негативно сказывается на качестве синтезируемой речи. Чтобы устранить потерю качества, для каждого дифона необходимо более детально учитывать контекст, что приводит к значительному увеличению размеров исходной базы данных. Использование полуслогов и более крупных элементов в качестве исходных имеет те же недостатки, что и использование дифонов.

Исходя из представленной выше информации, в данной работе в качестве исходных элементов для синтеза будут использоваться аллофоны.

1.5. Мини- и макси-наборы аллофонов для синтеза русской речи

Всего в русском языке насчитывается 42 фонемы, из них 6 гласных и 36 согласных. В потоке речи акустико-артикуляционные характеристики фонем изменяются, что и приводит к появлению аллофонов (оттенков фонем). Аллофоны делятся на два типа: позиционные и комбинаторные.

Позиционные аллофоны определяются положением фонемы по отношению к ударному слогу в слове [8]. Основными фонетическими коррелятами ударения в современном русском литературном языке служат длительность и спектральные характеристики гласных. По этим параметрам в СРЛЯ выделяется двухкомпонентное просодическое ядро слова, состоящее из ударного и 1-го предударного слогов [7]. Ударный и первый предударный слоги противопоставлены другим слогам. Русский лингвист А. А. Потебня описал ритмическую схему слова с просодическим ядром при помощи так называемой формулы Потебни: 112311. Здесь цифрой 3 обозначен ударный слог, цифрой 2 – первый предударный, цифрой 1 – остальные безударные слоги. Эта ритмическая схема является яркой особенностью русского языка и отличает его от всех других языков мира. Типичной схемой индоевропейских языков является схема 1213121.

Комбинаторные аллофоны определяются ближайшим контекстом фонемы [8], то есть являются результатом действия аккомодации (коартикуляции) и ассимиляции. Эффекты аккомодации и ассимиляции в большей степени заметны внутри слоговых комплексов, в меньшей степени на стыках слогов и в еще меньшей, но по-прежнему заметной, на стыках слов внутри синтагмы. На стыках синтагм эти эффекты практически не действуют. Таким образом, описать фонемы в потоке речи (аллофоны) можно с различной степенью детализации.

Речевой корпус должен удовлетворять следующим требованиям:

* быть фонетически полным, то есть содержать все основные варианты фонем;
* объем корпуса должен быть по возможности минимизирован;
* корпус должен быть фонетически сбалансированным.

Обычно выделяют и используют два практически обоснованных варианта набора аллофонов: мини- и макси-наборы. В макси-наборах выделяют пять позиционных аллофонов гласных (ударный, частично ударный, первый предударный, не первый предударный, заударный) и двадцать два комбинаторных аллофона гласных, итого для шести гласных русского языка выделяют 2700 аллофонов. Для согласных всего выделяют 1440 аллофонов [6]. В мини-наборе создается 240 аллофонов гласных и 180 аллофонов согласных [6]. Эти оценки являются теоретическими и сильно завышены, поскольку на практике многие аллофоны либо не используются, либо не являются смыслоразличающими, поэтому ими можно пренебречь. В итоге в макси-наборе практически используются 1550 аллофонов гласных и 209 аллофонов согласных, а в мини-наборе – 175 аллофонов гласных и 81 аллофон согласных.

Очевидно, что использование макси-набора представляет собой объемную и сложную задачу, поэтому в данной работе будет использоваться мини-набор. В связи с тем, что для записи базы диктора необходимо производить в студийных условиях, в качестве речевой базы будет использоваться корпус RUSLAN (Russian Spoken Language Corpus For Speech Synthesis) [11]. RUSLAN является открытым корпусом русского разговорного языка, содержит 22200 аудиозаписей с текстовыми аннотациями и является крупнейшим с точки зрения продолжительности речи одного диктора корпусом. Корпус распространяется под лицензией CC BY-NC-SA 4.0.

2. Конструкторский раздел

2.1. Требования к программе

Программа должна синтезировать речь по заранее неизвестному тексту на русском языке. Для синтеза используется метод аллофонного синтеза речи по фонетическим правилам. Входные данные подаются программе в виде файла. Синтезированная речь должна быть сохранена в виде аудио файла.

2.2. Описание входных и выходных данных и ограничений

Входными данными для программы является текстовый файл. Файл может содержать в себе:

* собственно текст на русском языке, состоящий из предложений или отдельных слов;
* аббревиатуры;
* сокращения;
* специальные символы (список допустимых символов приведен в );
* многоразрядные и дробные числа;
* телефонные номера;
* знаки пунктуации;
* разметка словесного ударения (подробнее см. ниже в этом разделе);
* слова с буквой «е» вместо «ё».

Файл не может содержать в себе (ограничения):

* иностранные слова;
* даты из римских цифр;
* обозначения времени и даты;
* интернет-адреса;
* какую-либо разметку текста (например, просодическую), кроме разметки ударений;
* математические выражения.

Программа позволяет пользователю размечать только словесное ударение ударения в словах. Для этого необходимо в слове после буквы, обозначающей ударный гласный, поставить знак «+». Любая другая разметка, такая как синтагматическое и фразовое ударение и т. д., программой не поддерживается.

Важным ограничением разрабатываемой программы является отсутствие полноценного просодического оформления текста. Программа делит текст на синтагмы опираясь только на знаки пунктуации (пунктуационные синтагмы), членение на лексические и синтаксические синтагмы не производится, потому что эта задача является объемной и сложной и выходит за рамки данной работы.

Функциональная модель (IDEF0-диаграмма) разрабатываемой системы представлена на рисунках *1* – *4*.

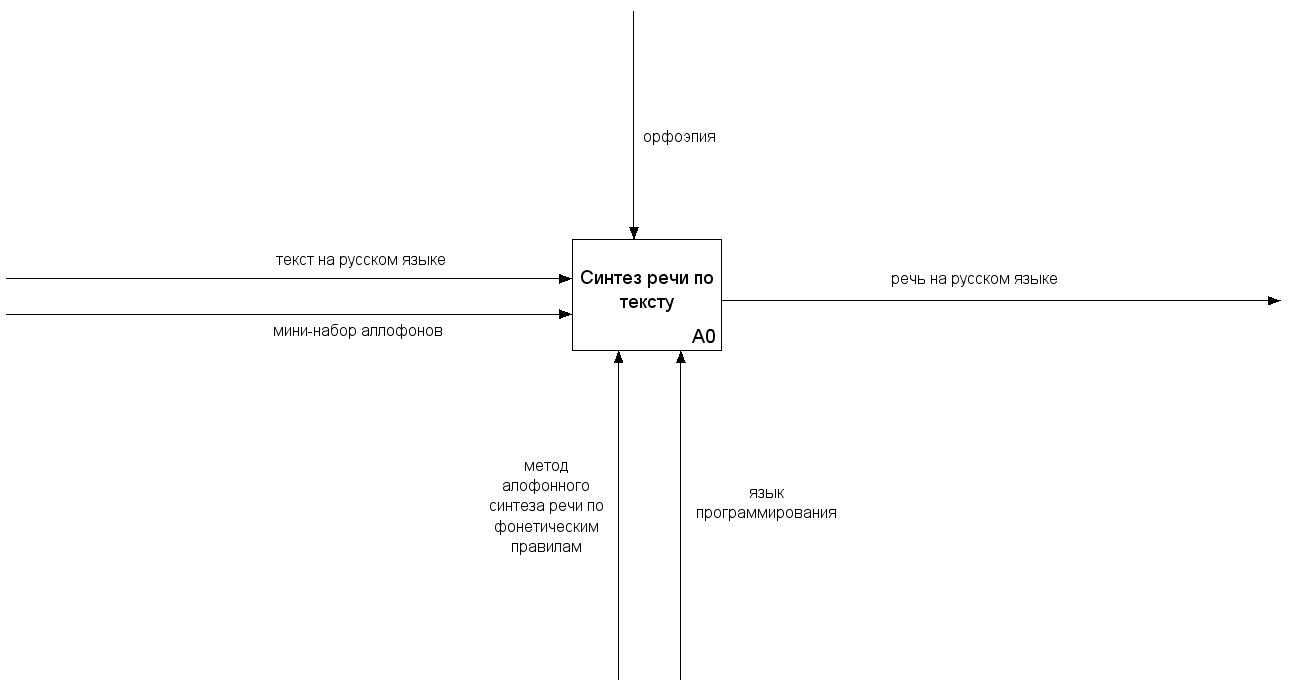


Рисунок 1. Контекстная диаграмма системы.

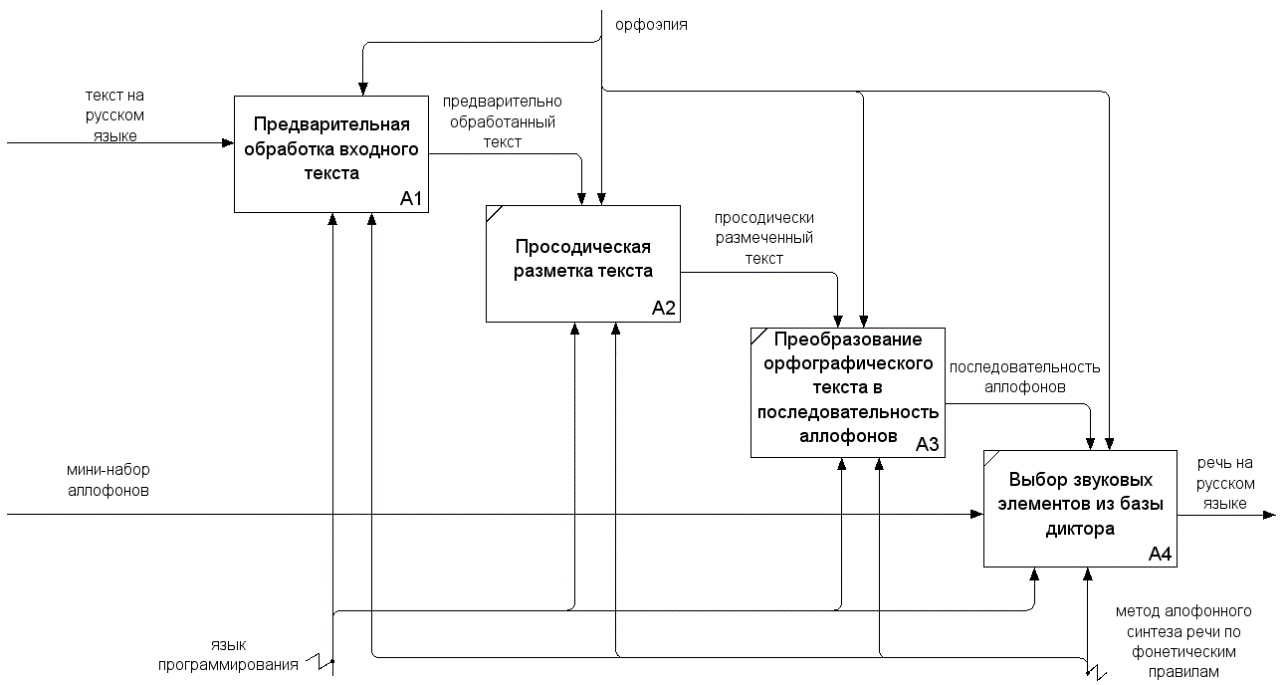


Рисунок 2. Декомпозиция работы «Синтез речи по тексту».

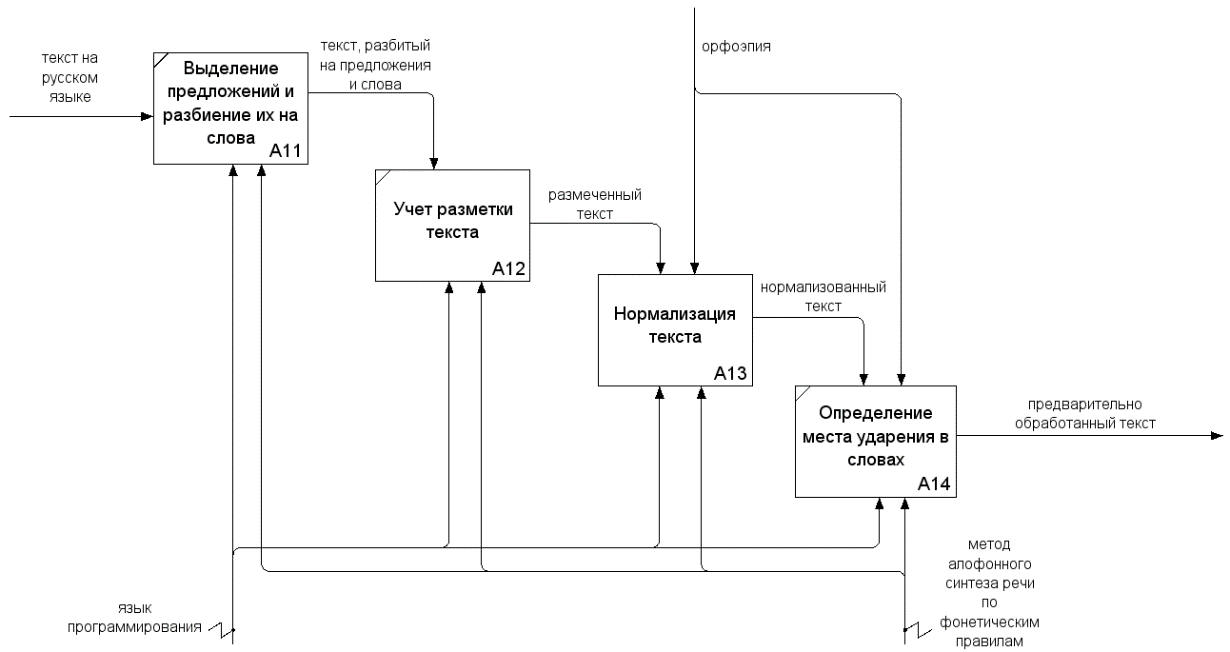


Рисунок 3. Декомпозиция работы «Предварительная обработка входного текста».

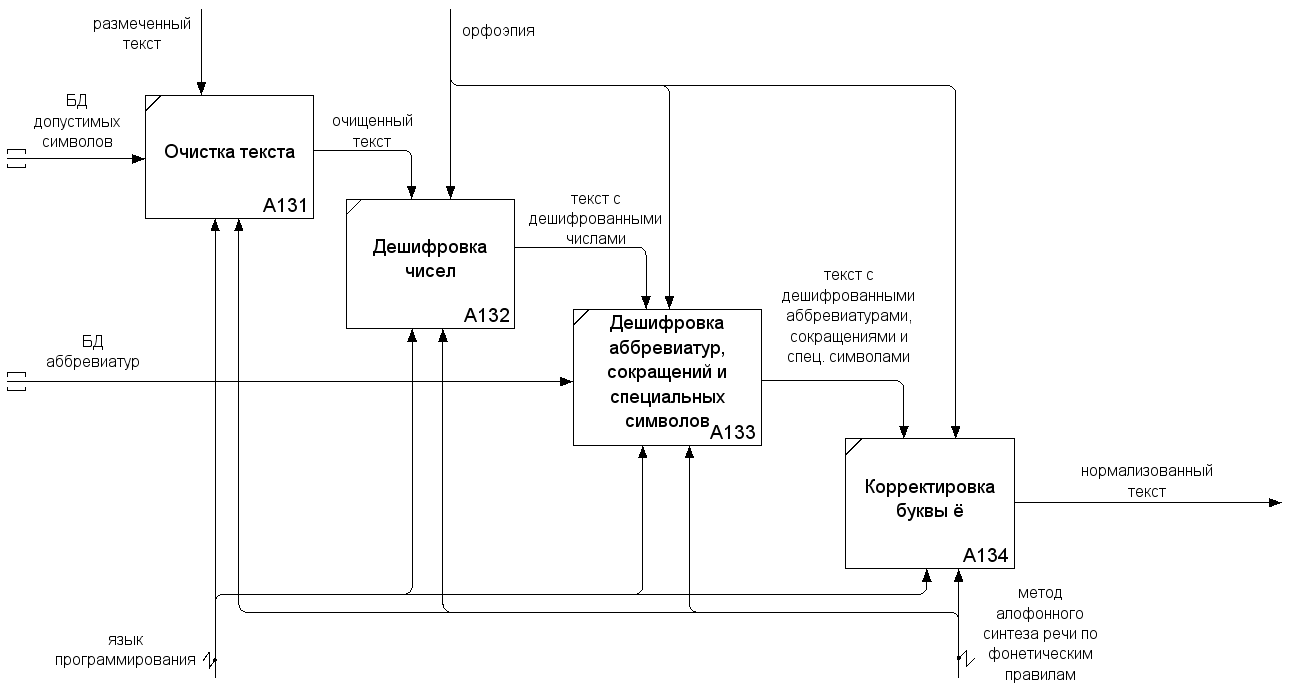


Рисунок 4. Декомпозиция работы «Нормализация текста».

Список литературы

1. Рыбин, С. В. Синтез речи. Учебное пособие по дисциплине «Синтез речи». / С. В. Рыбин. – СПб: Университет ИТМО, 2014. – 92 с.

2. Amazon Polly [Электронный ресурс]. – Режим доступа: <https://aws.amazon.com/ru/polly/>, свободный (дата обращения 29.01.2021).

3. Watson Text to Speech [Электронный ресурс]. – Режим доступа: <https://www.ibm.com/ru-ru/cloud/watson-text-to-speech>, свободный (дата обращения 29.01.2021).

4. Microsoft Azure. Преобразование текста в речь [Электронный ресурс]. – Режим доступа: https://azure.microsoft.com/ru-ru/services/cognitive-services/text-to-speech/, свободный (дата обращения 29.01.2021).

5. Cloud Text-to-Speech [Электронный ресурс]. – Режим доступа: <https://cloud.google.com/text-to-speech/>, свободный (дата обращения 29.01.2021).

6. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках [Электронный ресурс]. – Режим доступа: http://www.dialog-21.ru/digests/dialog2006/materials/html/lobanov.htm, свободный (дата обращения 31.01.2021).

7. Князев, С. В. Современный русский литературный язык: Фонетика, орфоэпия, графика и орфография: Учебное пособие для вузов. – 2-е изд., перераб. и доп. / С. В. Князев, С. К. Пожарицкая. – М.: Академический Проект; Гаудеамус, 2011. – 430 с.

8. Лобанов, Б. М. Компьютерный синтез и клонирование речи. / Б. М. Лобанов, Л. И. Цирульник. – Минск: «Белорусская наука», 2008.

9. Корпус устной русской речи [Электронный ресурс]. – Режим доступа: <https://cloud.google.com/text-to-speech/>, свободный (дата обращения 01.02.2021).

10. Huang, Xuedong. Spoken language processing: a guide to theory, algorithm, and system development. / Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. – Upper Saddle River, New Jersey: Prentice Hall PRT, 2001.

11. RUSLAN: Russian Spoken Language Corpus For Speech Synthesis [Электронный ресурс]. – Режим доступа: https://ruslan-corpus.github.io/, свободный (дата обращения 10.02.2021).