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

Для достижения поставленной цели необходимо решить следующие задачи:

• провести аналитический обзор существующих методов синтеза речи и выбрать подходящий;

• спроектировать архитектуру системы синтеза речи;

• выбрать технические средства для реализации спроектированного программного обеспечения;

• реализовать алгоритм по выбранному методу;

• провести тестирование разработанного ПО;

• оценить качество речи, синтезируемой с помощью разработанного ПО.

//

В современном мире существует необходимость создания систем разговорного языка. Такие системы сочетают в себе как распознавание речи, так и ее синтез. Успешное развитие этой технологии повысит доступность компьютеров и автоматизированных систем для широкого круга пользователей. 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.

//

В данной работе будет использоваться синтез по фонетическим правилам, потому как это единственный метод, позволяющий озвучить произвольный заранее неизвестный текст.

//

В данной работе будут использоваться аллофоны, потому что аллофоны являются наилучшим компромиссом между возможным качеством синтезируемой речи и объемом БД исходных элементов.

//

На слайде представлена функциональная модель реализации метода, на диаграмме видно, что система синтеза речи по тексту должна состоять из четырех подсистем: лингвистического текстового процессора, просодического процессора, фонетического процессора и акустического процессора.

//

Задачей лингвистического текстового процессора является предварительная обработка текста, которая включает в себя восемь этапов.

Выделение предложений.

Разбиение предложений на слова.

Очистка текста – это удаление неинформативных для синтеза речи символов.

Корректировка буквы «ё» необходима, так как правила русского языка не запрещают писать букву «е» вместо «ё». Однако в речи при произнесении «е» вместо «ё» слушающий сразу заметит неточность.

Для каждого слова необходимо определить положение словесного ударения. Для этого используется полная акцентуированная парадигма по А. А. Зализняку.

Фонетические исключения не подчиняются стандартным правилам преобразования буква-фонема, поэтому их необходимо заменить на эквиваленты, которые этим правилам подчиняются.

Формирование фонетических слов – это присоединение клитик к другим фонетическим словам. В русском языке существует ряд служебных слов, которые не имеют ударения и являются частью предыдущего (энклитики) или следующего (проклитики) фонетического слова.

//

Текст может содержать в себе:

• собственно текст на русском языке, состоящий из предложений или отдельных слов;

• аббревиатуры;

• знаки пунктуации;

• слова с буквой «е» вместо «ё».

Текст имеет ограничения и не может содержать в себе, например иностранные слова. Это связано с тем, что для их правильного озвучивания необходим синтаксический и семантический анализатор.

//

Под дешифровкой аббревиатур здесь понимается установление правил их чтения. Дешифровка аббревиатур необходима в связи с тем, что правила чтения аббревиатур в русском языке отличаются от правил чтения обычных слов. Чтобы чтение аббревиатуры было правильным, необходимо завести возможно более полную базу данных аббревиатур и их произношений. Если же в тексте встретилась аббревиатура, не присутствующая в базе данных, необходимо руководствоваться правилами чтения аббревиатур с учетом исключений.

//

Просодический процессор выделяет в каждом предложении последовательности слов, связанные синтаксической связью, которые представляют из себя цельные просодические единицы (синтагмы). В речи синтагмы отделяются друг от друга паузами.

В данной работе производится выделение только пунктуационных синтагм. Длинные участки текста, расположенные между знаками препинания могут звучать монотонно и осложнять восприятие речи, однако более глубокий анализ требует разработки серьезной системы семантического анализа.

Считается, что пунктуационные синтагмы ограничены знаками препинания, но после сочинительных союзов граница синтагмы не проводится.

//

Задачей фонетического процессора является преобразование орфографического текста в последовательность аллофонов. Фонетический процессор состоит из двух подсистем – блока преобразования текста в последовательность фонем (фонемный транскриптор) и блока преобразования фонемной транскрипции в последовательность аллофонов.

Фонемная транскрипция отражает современные правила произношения. Существуют правила преобразования орфографического текста в фонемную последовательность, а также целый ряд сложившихся исключений. Блок преобразования буква-фонема работает на этих правилах.

В потоке речи в различных условиях акустико-артикуляционные характеристики фонем изменяются. Таким образом появляются оттенки фонем, аллофоны. Аллофоны разделяют на позиционные, определяющиеся положением фонемы по отношению к словесному ударению, и комбинаторные, определяющиеся ближайшим контекстом фонемы и обусловленные ассимиляцией и коартикуляцией. Преобразованием фонем в аллофоны занимается блок преобразования фонема-аллофон.

//

Задачей акустического процессора является выбор необходимых элементов из базы данных исходных элементов и их склейка в общий выходной аудиофайл. Между фонетическими словами добавляется короткая пауза, а между синтагмами добавляется более длинная пауза. После склейки всех аллофонов на синтагму накладывается эффект fade in, чтобы синтезированная речь звучала чуть более эмоционально. Fade in – это постепенное увеличение громкости аудио-дорожки к ее концу.

Для того, чтобы синтезированная речь звучала более равномерно, необходимо сгладить места склейки. Для этого используют алгоритм интерполяции звуковых сигналов (из библиотеки pydub) – увеличения частоты дискретизации сигнала в N раз. Непосредственно интерполяция выполняется библиотекой.

//

Для поддержки разработанной системы синтеза речи необходимо пять хранилищ данных: четыре словаря и БД исходных элементов синтеза.

Для хранения словарей используется NoSQL база данных типа «ключ-значение». Всего создано 4 базы данных: БД ударений, слов фонетических исключений, согласных и аббревиатур.

Исходные элементы синтеза хранятся в файловой системе в виде звуковых файлов. Схема файлового хранилища представлена на рисунке.

//

Всего в русском языке насчитывается 42 фонемы. В потоке речи фонемы преобразуются в аллофоны (оттенков фонем). Исходные элементы синтеза – это аллофоны.

Обычно выделяют и используют два практически обоснованных варианта набора аллофонов: мини- и макси-наборы. В макси-наборе выделяют 4140 аллофонов, а в мини-наборе – 420 аллофонов.

Очевидно, что использование макси-набора представляет собой объемную и сложную задачу, поэтому в данной работе будет использоваться мини-набор.

В качестве речевой базы был использован открытый корпус русского разговорного языка RUSLAN.

//

Для тестирования разработанного программного обеспечения были разработаны unit-тесты для шести классов: класс преобразования буква-фонема, класс преобразования фонемаз-аллофон, класс деления последовательности аллофонов на открытые слоги, а также три класса для деления последовательности аллофонов на слоговые комплексы трех типов.

Все тесты были успешно пройдены.

//

Использование базового мини-набора аллофонов качество синтезируемой речи оказалось недостаточно высоким. Это объясняется, в первую очередь, тем, что взаимовлияние соседних аллофонов друг на друга в некоторых случаях оказалось настолько сильным, что при создании БД аллофонов не всегда удавалось провести между ними четкую границу. На рисунке приведена осциллограмма и спектрограмма слова «загремела». Границы первых трех звуков прослеживаются достаточно отчетливо, в то время как границы последующие звуков провести очень трудно.

Для того, чтобы преодолеть этот недостаток, можно в качестве исходных элементов синтеза были использованы аллослоги трех типов (в зависимости от сложности сегментации).

//

При оценке качества речи было использовано два критерия:

• разборчивость (фразовая, смысловая и словесная);

• естественность (субъективная характеристика).

Для исследования качеств синтезируемой речи был приглашен независимый слушатель, которому предлагалось прослушать контрольные примеры.

Для исследования фразовой и смысловой разборчивости слушателю было предложено прослушать семь контрольных фраз. При использовании только аллофонов слушатель смог понять только четыре контрольные фраз. Такую речь тяжело понимать и приходится из контекста догадываться, что было сказано. Естественность речи была оценена как очень низкая, так как имеет место большое количество мест «склейки» исходных элементов синтеза. При использовании слоговых комплексов слушатель также понял все семь контрольных фраз, однако речь стало намного легче воспринимать и отпала необходимость «достраивать» фразу из контекста. Естественность речи повысилась, однако назвать ее натуральной (имеющей высокую естественность) все равно нельзя.

Для исследования словесной разборчивости слушателю было предложено прослушать семь слов, не часто встречающихся в языке. При использовании в синтезе только аллофонов неподготовленному слушателю удалось опознать лишь два из семи слов, что говорит об очень низкой разборчивости речи. При использовании слоговых комплексов любых типов слушатель смог опознать все слова. Если говорить об естественности, то в первом случае она субъективно оценивается скорее как низкая, а во втором – как средняя.

Что касается объема базы данных исходных элементов, то при аллофонном синтеза с использованием мини-набора акустическая БД должна содержать 327 элементов синтеза. При использовании аллослогов третьего типа (легко сегментируемых) необходимо 386 элементов. При этом при подготовке БД аллослогов были созданы только слоги, которые содержатся в контрольных примерах. При этом при подготовке БД аллослогов были созданы только слоги, которые содержатся в контрольных примерах.

//

В результате выполненной работы были решены все поставленные задачи.

Была достигнута поставленная цель: разработан инструмент, позволяющий синтезировать речь по заранее неизвестному тексту на русском языке. Разработанное программное обеспечение было протестировано и прошло все тесты.

//

В качестве дальнейшего развития проекта можно предложить следующее:

• убрать ограничения на входной текст, а именно, добавить возможность озвучивать иностранные слова, специальные символы, многоразрядные и дробные числа, телефонные номера, сокращения, даты из римских цифр, обозначения времени и даты, интернет-адреса, математические выражения;

• добавить поддержку пользовательской разметки (ударения и просодическая разметка);

• разработать или интегрировать синтаксический и семантический анализатор для более качественного деления текста на синтагмы и для возможности озвучивать ограничения, приведенные в первом пункте;

• добавить интонационное оформление текста.

//

Спасибо за внимание! Готова ответить на ваши вопросы.