МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение  
высшего образования

**"Южно-Уральский государственный университет**

**(национальный исследовательский университет)"**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

**ОТЧЕТ**

**о выполнении научно-исследовательской работы**

бакалавра направления 02.03.02 «Фундаментальная информатика и информационные технологии»

Автор работы:

студент группы КЭ-201

Сафина Алина Маратовна С

(Фамилия Имя Отчество студента)

Научный руководитель:

Доктор физ. мат. наук, доцент

<ученая степень, ученое звание>

М. Л. Цымблер

<И. О. Фамилия руководителя>

Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_

Дата: \_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись: \_\_\_\_\_\_\_\_\_\_\_\_\_

Челябинск 2020

ЧЧЧЧ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение  
высшего образования

**"Южно-Уральский государственный университет**

**(национальный исследовательский университет)"**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

УТВЕРЖДАЮ

Зав. кафедрой СП

\_\_\_\_\_\_\_\_\_\_ Л.Б. Соколинский

01.03.2021 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

студенту группы КЭ-201

Сафина Алина Маратовна С

(Фамилия Имя Отчество студента)

обучающемуся по направлению 02.03.02 «Фундаментальная информатика и информационные технологии»

1. **Тема НИР в семестре**«Применение технологий искусственного интеллекта для синтеза речи на примере разработки простого синтезатора речи »
2. **Срок сдачи студентом законченной работы** 31.05.2021 г.
3. **Перечень подлежащих разработке вопросов**
4. Обзор научных статей по тематике применения искусственных нейронных сетей для синтеза речи
5. Разработка простого синтезатора речи

**Научный руководитель**

Доцент кафедры СП,

Доктор физ. мат. наук М.Л. Цымблер

**Задание принял к исполнению** А.М. Сафина

Оглавление

[1. ОБЗОР НАУЧНЫХ СТАТЕЙ ПО ТЕМЕ «СИНТЕЗ РЕЧИ» 6](#_Toc73278484)

[1.1 История зарождения 6](#_Toc73278485)

[1.2 Основные виды и подходы 9](#_Toc73278486)

[1.3 Применение технологии синтеза речи 13](#_Toc73278487)

[1.4 Проблемы и особенности 14](#_Toc73278488)

[2. ПРОЕКТИРОВАНИЕ 17](#_Toc73278489)

[2.1 Архитектура программы 17](#_Toc73278490)

[2.2 Модульная структура программы 18](#_Toc73278491)

[3. РЕАЛИЗАЦИЯ 20](#_Toc73278492)

[3.1 Реализация программы 20](#_Toc73278493)

[3.2 Главный модуль генерации речи 20](#_Toc73278494)

[3.3 Модуль анализа текста 20](#_Toc73278495)

[3.4 Речевая база 20](#_Toc73278496)

[3.5 Модуль выбора звуковых секций 20](#_Toc73278497)

[3.6 Модуль генерации речи 21](#_Toc73278498)

[4. ТЕСТИРОВАНИЕ 22](#_Toc73278499)

[**4.1. Основной модуль генерации речи** 22](#_Toc73278500)

[**4.2. Анализ текста** 22](#_Toc73278501)

[**4.3. Выбор звуковых секций** 23](#_Toc73278502)

[**4.4. Генерация речи** 24](#_Toc73278503)

[5. РАБОТА ИТОГОВОГО ПРИЛОЖЕНИЯ 26](#_Toc73278504)

[ЗАКЛЮЧЕНИЕ 27](#_Toc73278505)

[Список литературы: 28](#_Toc73278506)

[Приложение 1 31](#_Toc73278507)

[Приложение 2 31](#_Toc73278508)

[Приложение 3 31](#_Toc73278510)

[Приложение 4 32](#_Toc73278512)

**ВВЕДЕНИЕ**

Синтез речи – процесс восстановления формы речевого сигнала по его параметрам [5]. Зарождение данной технологии произошло еще во 2-й половине XVIII века [2].

В начале это был аппарат, состоящий из резонансных трубок [1], подключаемых к органу. Данная технология развилась, и сейчас мы имеем синтезаторы речи, которые могут за довольно короткий промежуток времени конвертировать текст в голос, соблюдая, кроме того, знаки препинания и учитывая интонацию.

Технология синтеза речи имеет применение в различных областях, например она может быть использована в информационно-справочных системах, для помощи людям с нарушениями зрения [5] (пример – программа [JAWS](https://www.freedomscientific.com/products/software/jaws/)), в робототехнике, при создании разнообразных ботов и голосовых помощников (пример – [Siri](https://www.apple.com/siri/)) и как предмет искусства (существуют так называемые «Вокалоиды» [16], которые имитируют голос поющего человека, существуют «электронные» знаменитости, такие как [Хатсунэ Мику](https://ec.crypton.co.jp/pages/prod/virtualsinger/cv01_us).

Для реализации данной технологии сейчас активно используется искусственный интеллект, в частности, популярные в настоящее время нейронные сети (пример – [Vera Voice](https://veravoice.ai/), WaveNet [17]).

В данном отчете был проведен анализ научных публикаций по тематике «Синтез речи», был написан их обзор, с целью лучшего понимания выбранной темы, в том числе истории зарождения данной технологии, видов и методов ее реализации, применения, проблем и особенностей.

**1. ОБЗОР НАУЧНЫХ СТАТЕЙ ПО ТЕМЕ «СИНТЕЗ РЕЧИ»**

1.1 История зарождения

Первые синтезаторы появились во 2-й половине XVIII в. и насчитывают более чем 200-летнюю историю [1, 2, 4]. Это были механические устройства, которые требовали участия обученного оператора-исполнителя [1]. В то время было создано множество примечательных устройств, например, стоит отметить механический синтезатор Р. Риша, который по форме практически точно копировал речевой тракт человека [1]. В то время это считалось самым очевидным способом воссоздать естественную речь.

В 1779 г. Петербургской Академией наук была объявлена ежегодная премия за создание устройства, способного воспроизводить пять русских гласных звуков. Лучшим образом эту задачу выполнил немецкий ученый Христиан Готлиб Кратценштейн (1723-1795). Он создал систему резонаторов, которая при помощи воздушного потока воспроизводила русские гласные [1, 4]. Первые синтезаторы могли воспроизводить лишь гласные звуки, однако чуть позже (в 1791 г.) австрийский ученый Вольфганг фон Кемпелен представил аппарат, способный производить и согласные звуки, а также некоторые короткие фразы. В то время управление синтезатором речи можно было сравнить с игрой на музыкальном инструменте [1, 2, 4].

Далее, в течение ста лет, большого скачка в развитии данной технологии не наблюдалось.

В ХХ в. Начались попытки создать электронные синтезаторы речи. Несмотря на это, разработка механических синтезаторов проводилась до 60-х годов. Это было связано с дефицитом сложных электрических компонентов [1] и сложностью вычислений.

Было установлено, что гласные звуки различаются резонансными частотами, названными впоследствии формантами [2, 4]. Первый электрический синтезатор речи был создан в 1922 г. Дж. Стюартом. Схема включала в себя электрический зуммер: он моделировал работу голосовых связок, а также два индуктивно-емкостных резонатора, которые моделировали резонансы горла и ротовой полости. Опять же, данное устройство могло синтезировать только гласные звуки [1, 4].

Следующей ступенью развития стало появление вокодеров (система кодирования и декодирования речи, в ней используются различные методы сжатия полосы частот для передачи сигналов) [1, 4]. «Водер» (VODER), созданный Гомером Дадли, Р. Ришем и С. Уоткинсом был продемонстрирован на Всемирной выставке в Нью-Йорке в 1939 г. Он был способен генерировать фрагменты довольно качественной связной речи с высоким уровнем разборчивости [2, 4].

Важным этапом в развитии синтеза речи стала разработка звукового спектрографа в 1946 г. [1, 4].

Вместо прямого моделирования речевого тракта человека, основным методом создания синтезатора стало моделирование акустических характеристик речевого сигнала [1, 4].

Важной вехой в истории синтеза речи стало развитие акустической теории речеобразования в 1960 г., которая создала теоретическую базу для создания артикуляционных, формантных и использующих линейное предсказание синтезаторов основанных на ней. Эти три метода называют также технологиями синтеза первого поколения [1, 4].

Подробнее с основными подходами к разработке синтезатора речи можно ознакомиться в разделе «Основные виды и подходы» данной работы.

В 1962 г. Джон Л. Келли впервые в мире применил для синтеза речи компьютер. Это был IBM 704 [2].

В связи со стремительным развитием компьютерной техники в середине 60-х годов, перед разработчиками автоматического синтеза речи встала потребность реализации более сложной задачи озвучивания любого сообщения, вводимого в компьютер в текстовом виде и неизвестного системе синтеза заранее. Это привело к развитию синтезаторов типа «Текст–Речь» (Text-to-Speech или сокращенно TTS) [1, 4]. В таких синтезаторах появился блок лингвистической обработки, который не зависел от акустического блока, в отличии от ранних синтезаторов [1]. Первая полноценная система TTS была создана в 1968 г. В Японии Норико Умеда и его коллегами. Она была основана на артикуляционной модели акустического блока. Анализ текста и расстановка пауз производились при помощи сложных правил. Речь, производимая этой системой, была разборчивой, но довольно монотонной. В дальнейшем алгоритмы лингвистической предобработки текста усложнялись благодаря увеличению скорости компьютерного анализа данных и объема памяти для хранения вспомогательной информации (словарей, речевых баз, моделей и т. п.) [1, 4]. Благодаря этому появилась возможность более точно представлять необходимые для акустического синтеза детальные фонетические описания, а именно фонетическую транскрипцию и просодические характеристики сегментных единиц (длительность, частота основного тона и громкость), получаемые на основе интонационных моделей [4].

К третьему поколению технологий автоматического синтеза речи обычно относят селективный синтез речи и синтез на основе скрытых Марковских моделей (HMM – hidden Markov models. В англоязычных источниках метод называют «**Unit selection»**. Их общей чертой является использование для автоматического синтеза речи больших объёмов речевых данных, а также высокая естественность и качество синтезированной речи [1, 4]. Описание данных методов можно найти в следующем разделе.

В институте Современных Телекоммуникационных Исследований (Advanced Telecommunications Research) в Японии в конце 1980-х — начале 1990-х годов, где Иосинори Сагисака использовал обширную базу данных, хранившую множество различных речевых контекстов для каждого дифона [2]. **Дифон** – отрезок, начинающийся в середине одного звука и заканчивающий в середине следующего [1].

В конце 80-ых годов ХХ в. для синтеза речи впервые были применены технологии нейронных сетей, однако успехов это не принесло из-за огромной на то время стоимости вычислений [2].

С 2010 года глубокие нейронные сети активно применяются для задач статистического параметрического синтеза речи. Существуют различные методы работы с ними, в том числе применение LSTM, генеративных состязательных сетей, END-TO-END моделей [2].

Синтез эмоциональной и экспрессивной речи тоже стал важной темой для современных исследований [1], ведь для повышения качества «машинной» речи это необходимо. Для этой задачи были применены системы с явным контролем, формантный синтез, конкатенации дифонов, корпусный подход, комбинированная система и тд. [2], описания приведены в следующем разделе.

1.2 Основные виды и подходы

**Артикуляционный синтез. Первое появление – 1950 г., первый синтезатор – DAVO (Dynamic Analog of the Vocal tract)**[1, 4].В нем речевой сигнал синтезируется на основе моделирования процесса речеобразования с учетом артикуляции, количественной оценки формы речевого тракта, его резонансных свойств и характеристик звуковых источников. Из-за сложности подобного моделирования, в данный момент он используется для научных исследований в области фонетики и физиологии речи, а в системах синтеза используются более простые подходы [1, 4]. Также данный вид синтеза речи напрямую связан с конструированием модели головы или лица в процессе говорения (аудиовизуальный синтез).

**Формантный синтез. Первое появление – 1953 г., первый синтезатор – PAT (Parametric Artificial Talker)**[1, 4].Состоял из трех электронных формантных резонаторов, соединенных параллельно, на вход подавался гармонический сигнал или шум, управлялся шестью временными функциями. Один из главных плюсов – простота управления. Минус – артефакты речи из-за того, что произносимый текст «склеивался» из различных фрагментов [2]. Существовали формантные параллельные и каскадные синтезаторы [1, 4].

**Каскадный синтезатор:** выходной сигнал каждого резонатора является входным сигналом следующего [1, 4]. Формантные резонаторы соединены последовательно [4].

**Параллельный синтезатор:** амплитуда каждого формантного резонатора контролируется отдельно [1, 4].

Первым формантным синтезатором речи для русского языка стал «ФОНЕМОН-1», появившийся в начале 70-х годов в Минске [2, 4].  
 **Линейное предсказание. Первое появление – середина 60-х годов, впервые применена в устройствах типа TI Speak’n’Spell (1980 г.)**[1, 4].Данный метод позволяет напрямую использовать параметры передаточной функции речевого тракта. Является альтернативой формантному синтезу [1, 4]. Основной частью данного синтеза был коэффициент линейного предсказания, принцип работы заключался в том, что необходимые единицы синтеза, например слова, оцифровывались и переносились в память, а затем воспроизводились в определенном порядке. Обычно они требовали строгого описания произносимого текста. Данная технология поспособствовала развитию конкатенативного синтеза [1, 4].

В это же время началось развитие синтезаторов типа «Текст-Речь» (Text-to-Speech).

**Конкатенативный синтез.** Благодаря большему объему доступной памяти в памяти компьютера могли храниться фрагменты реальной речи человека и путем склейки создавалась основа акустического сигнала [1, 4]. Данные фрагменты речи были довольно короткими и предложения составлялись путем склеивания из-за чего этот вид синтеза часто выдавал неестественную речь, кроме того проблемой данного метода является недостаточная гибкость в изменении тембра голоса [1, 2].

Наиболее часто используется алгоритм PSOLA (Pitch Synchronous Overlap and Add), который был разработан в 1985 г. [1]. Более подробное его описание можно найти в разделе о естественности речевого сигнала.

**Корпусный подход. Конец 1980-х - начало 1990-х годов.** В данном подходе была использована обширная база данных речевых контекстов для каждого дифона. Для поиска лучшей комбинации дифонов использовалась функция акустической дистанции. Данный подход показал, что при достаточном кол-ве дифонов и их правильной комбинации на выходе может получиться качественная речь, близкая к естественной [2].

**Селективный синтез речи (Unit Selection). Появление – 1992 г., первая система n-Talk и CHATR**[1, 4].Доминирующая в данный момент технология. Основой для этого подхода является конкатенативный синтез. При селективном синтезе для каждой единицы синтеза производится выбор наиболее подходящего варианта, взятого из озвученных предложений естественного языка [1]. В отличии от конкатенативного метода программа обучается на имеющихся речевых данных и получает модель соответствия характеристик речи нужным физическим параметрам звуковых единиц. Данный метод занимает меньший объем памяти и позволяет получить наиболее близкую по характеристикам к естественной речь, из-за чего является доминирующей технологией[1, 4]. При селективном синтезе для каждой единицы синтеза производится выбор наиболее подходящего варианта, взятого из озвученных предложений естественного языка [1, 4]. Для определения того, насколько элемент базы подходит для синтеза нужной единицы, используются функции **стоимости замены** и **стоимости связи**[2].

**Статистический параметрический синтез. Впервые предложен в 2000 г**. [2].В отличии от конкатенативного метода программа обучается на имеющихся речевых данных и получает модель соответствия характеристик речи нужным физическим параметрам звуковых единиц, т.е генерирует «среднее» [1, 2], что схоже с принципом работы нейронных сетей. Наиболее распространенной техникой в данном подходе является метод, основанный на использовании СММ (скрытных Марковских моделей) [1]. Исследователи отмечают преимущество данного подхода над другими благодаря его гибкости, из-за чего можно меня тембр и эмоциональную окраску речи [2]. Данный метод в настоящее время самый широко используемый [10].

Речевые синтезаторы делят на два основных типа: с **ограниченной** (речевая база произносится диктором заранее) и **неограниченной** словарной базой (элементами речи являются фонемы или слоги, которые затем соединяются по фонетическим правилам) [4].

**Синтез речи по фонетическим правилам**. Различают следующие виды данного синтеза [4]:

* Микросегментный
* Аллофонный
* Дифонный
* Полуслоговый
* Слоговый
* Синтез из различных единиц произвольного размера.

Они различаются размером исходных элементов.

Прогресс в аппаратных средствах и программном обеспечении позволил ученым использовать нейронные сети для улучшения качества синтетической речи [2].

Глубокие нейронные сети успешно применяются в данной области с 2010 г. [2].

Применения глубоких нейронных сетей были разделены на три подхода [18]:

1. Для обучения акустической модели и каждого акустического кластера отдельно.
2. Для предсказывания акустических параметров, со входными данными в виде лингвистических представлений. Входные данные моделируются с помощью совместного распределения вероятностей.
3. Для предсказывания акустических параметров, со входными данными в виде лингвистических представлений. Входные данные моделируются с помощью условного распределения.

**LTSM (Long-Short Term Memory)**, представленный в 2015 г. Хейга Сеном и Хасим Сак, является синтезом речи на основе нейронных сетей долгой краткосрочной памяти. Он используется для предсказывания акустических параметров, длительности пауз, фонем. В настоящее время применение LTSM является классическим подходом. Существует синтезатор речи на основе LTSM для мобильных устройств [2].

**Генеративные состязательные сети (Generative adversarial network – GAN)** успешнее решают проблему сглаженности речевого сигнала. Состоят из двух конкурирующих нейронных сетей, условно разделяемых на генератор G (генерирует акустический вектор из лингвистического представления) и дискриминатор D (учится определять, какие акустические параметры получены из реального сигнала, а какие нет). Используются в том числе для прогнозирования эмоций речевого сигнала [2].

Несмотря на огромное количество существующих в настоящий момент подходов, некоторые из которых еще будут представлены далее, довольно трудно определить какой из них является наиболее подходящим в каждой конкретной ситуации. Это зависит от множества факторов, например различия языков, особенностей произношения, интонации, стресса, тембра и так далее  [13], о чем и повествует раздел «Особенности технологии».

1.3 Применение технологии синтеза речи

Данная технология имеет широкое распространение в наше время. Одной из задач синтеза речи является упрощение восприятия информации человеком.

Например, синтез речи используется для того, чтобы упростить людям с проблемами зрения восприятие текстовой информации. TextToSpeech Robot (TTSR) для данной цели был рассмотрен в этой статье [12]. Эта программа распознает введенный текст или уже готовое текстовое поле и конвертирует его в речь. Также она имеет функцию поддержки интернет-серфинга, позволяя читать текст прямо с веб-страницы. Пользователю всего лишь нужно выделить курсором нужный отрывок текста, а затем нажать на кнопку «Play». Озвученный текст можно сохранить, чтобы проиграть позже [12]. Внутренняя структура данной программы не отличается ничем особенным, главная ее проблема заключается в том, чтобы иметь максимально удобный интерфейс, учитывая, что ей будут пользоваться люди с проблемами зрения, а также предоставить высокое качество синтезированной речи, о чем повествует раздел «Особенности и проблемы синтеза речи» данной работы.

В данной статье [15] авторы задаются целью создать синтезатор речи для автоматического произношения, чтобы помочь не не-носителям языка корректно произносить имена и фамилии их родной страны [15]. Программа принимает на вход текст и предполагает к какому языку/диалекту он принадлежит. Далее программа озвучивает текст, используя определенный ранее язык как основной [15].

Синтез речи также имеет применение в области компьютерных игр. Автор данной статьи [8], представляет модель игры, где нужно управлять виртуальным агентом и решать головоломки. В зависимости от успеха или провала какого-либо действия, а также как просто реакция на происходящее, агент должен выражать свои эмоции посредством речи [8]. Применение данной технологии в этой области поможет игрокам более глубоко погружаться в игровой процесс и сопереживать герою, так как создается иллюзия его реальности. Более подробно об эмоциональности синтезированной речи в следующем разделе.

Когда встает задача о разработке синтезатора речи для какого-то конкретного языка, в процессе его создания изучается множество особенностей данного языка, например фонетических систем, пунктуации и т.д. Поэтому технология синтеза речи может применяться и для изучения лингвистики. В пример можно привести данную статью [9], авторы которой создать фонетико-акустическую базу данных для многоязычного синтеза речи по тексту на славянских языках, таких как белорусский, польский, русский.

Существует еще множество применений данной технологии, я упомянула лишь некоторые из них.

1.4 Проблемы и особенности

Одной из главных проблем данной технологии является естественность речевого сигнала. В основном она зависит от объема и покрытия речевой базы [3, 6], так как это дает возможность выбрать наиболее подходящий вариант и избежать речевых артефактов.

Модификация тона – это приближение синтезированной фразы к реально произнесенной. Она выполняется с помощью распределения полученных взвешенных фреймов на новые значения частоты, представляющей собой множество расстояний между окнами, им соответствующими. Существует множество алгоритмов, реализующих эту технологию, например TD-PSOLA, SPECINT, LP-PSOLA и т.д. [3].

При разработке синтезатора речи стоит учитывать множество нюансов, которые непосредственно влияют на качество речи. В данной статье [7], авторы выяснили, какие характеристики являются определяющими в создании звуков речи (фонем). Некоторые из них: относительная энергетика спектрального воздействия в определенных зонах, количество зон относительного спектрального воздействия, наклон частоты среза спектрального воздействия, добротность воздействия сигнала и т.д. [7]. Также для каждого языка существуют свои особенности произношения, интонации, грамматики и др. Например, в русском языке это склонения слов, такие части речи, как деепричастия и т.д. Более подробно особенности синтеза русской речи рассматриваются в данной статье [14].

Важной областью исследования технологии синтеза речи, о которой стоит упомянуть, является синтез эмоциональной и выразительной речи. Различия между нейтральной и выразительной проявляются в интонации, ритмике, паузах, ударениях и темпе произнесения [6]. Теоретически можно выделить три способа синтезировать речь в различных стилях (с помощью метода Unit Selection) [6]:

1. Создать несколько речевых баз с помощью одного голоса, но в разных стилях (интонациях).

2. Создать метки, соответствующие стилям, расставить их в базе данных и учитывать при синтезе.

3. Разработать просодические модели, настраиваемые на голос и стиль, учитывать их при выборе речевых элементов.

Модель экспрессивной речи, представленная в данной научной работе [11], состоит из двух частей: кодировщик эмоций и End-to-end Text-to-Speech модель, основывающаяся на нейронной сети. Представлен метод трансплантации эмоций и результаты тестирования.

**2. ПРОЕКТИРОВАНИЕ**

2.1 Архитектура программы

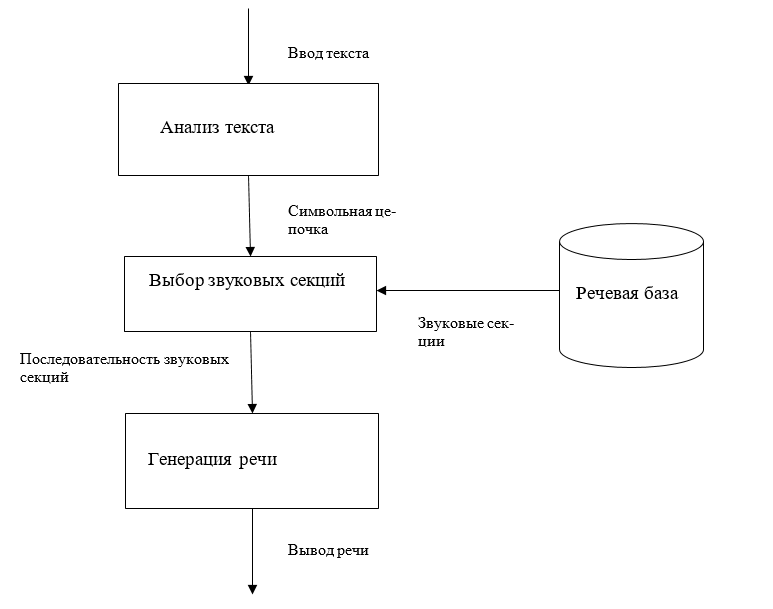
****

Рис. 1. Архитектура программы

В архитектуре программы (см. выше) представлены основные функции, которые должен выполнять готовый синтезатор речи. Программа представляет собой трехэтапный процесс, состоящий из анализа нужного текста, выбора звуковых секций из речевой базы и финального этапа: генерации речи в которой реализуется технология нейронных сетей. Программа работает через командную строку.

2.2 Модульная структура программы

main\_speech\_generation

text\_analys

speech\_generation

sound\_choose

Рис. 2. Модульная структура программы

**Описание модульной структуры:**

При запуске программы, модульная структура которой показана на рисунке 2 (см. выше) производится считывание текста из текстового файла для последующего синтеза, оно реализуется в главном модуле генерации речи *main\_speech\_generation*. Так же этот модуль вызывает все последующие и производит создание аудио-файла, который сохраняется на компьютере в конце работы программы. Подробное описание модуля *main\_speech\_generation* и его подмодулей представлено в разделе реализация, интерфейс модуля представлен в приложении 1 на рисунке 4 (см. ниже).

Модуль анализа текста *text\_analys* преобразует текст в цепочку слов, для поиска их в речевой базе. Подробное описание модуля *main\_speech\_generation* и его подмодулей представлено в разделе реализация, интерфейс модуля представлен в приложении 2 на рисунке 5 (см. ниже).

Затем модуль выбора звуковых секций *sound\_choose* выбирает нужные звуки из речевой базы и выстраивает их в нужном порядке. Подробное описание модуля представлено в разделе реализация, интерфейс модуля представлен в приложении 3 на рисунке 6 (см. ниже).

Далее функция генерации речи *speech\_generation* посредством использования нейронной сети и модуля pyttsx 3 склеиваетзвуковые секции воедино, устраняет речевые артефакты и подает цельную звуковую дорожку на вход в главный модуль генерации речи. Нейронная сеть в данной реализации обучается пользователем: она меняет параметры синтезатора речи, такие как скорость, громкость и т.д., встроенные в модуль pyttsx 3. Нейронная сеть обучается с помощью ввода 0 (плохо) и 1(хорошо) с клавиатуры. Подробное описание модуля представлено в разделе реализация, интерфейс модуля представлен в приложении 4 на рисунке 7 (см. ниже).

**3. РЕАЛИЗАЦИЯ**

3.1 Реализация программы

Можно выделить следующие этапы в реализации программы.

1. Работа с исходным текстом. Разделение его на отдельные части для выборки из речевой базы подходящих фрагментов.
2. Реализация голосового движка.
3. Работа с нейронной сетью для устранения речевых артефактов.
4. Реализация создания аудиофайла с полученной речью

3.2 Главный модуль генерации речи

Для считывания текста из файла использовался стандартный метод открытия файла и записи в переменную по строкам. Чтобы кириллица отображалась корректно был использован метод для считывания из файла с помощью стандарта кодирования utf-8.

3.3 Модуль анализа текста

Для разбиения текста на слова была реализована простая функция с циклом for, определяющая пробел как разделения слов. Результатом выполнения программы является список слов.

3.4 Речевая база

Речевая база реализована с помощью готовой базы RH Voice, находящейся в свободном доступе в интернете. Она автоматически заносит себя в доступную базу голосов модуля pyttsx 3 при установке.

3.5 Модуль выбора звуковых секций

Выбор звуковых секций происходит автоматически и реализован с помощью модуля pyttsx 3. С помощью инициализации переменной engine и использования встроенной функции pyttsx3.init() инициализируется звуковой движок, к которому можно обратиться с помощью функции engine.getProperty(), чтобы выбрать нужный голос из базы. Так же с помощью данной функции реализуется изменение свойств речи. Функция engine.say() воспроизводит текст внутри указанной в скобках переменной.

3.6 Модуль генерации речи

Выбор звуковых секций происходит автоматически и реализован с помощью модуля pyttsx 3. С помощью инициализации переменной engine и использования встроенной функции pyttsx3.init() инициализируется звуковой движок, к которому можно обратиться с помощью функции engine.getProperty(), чтобы выбрать нужный голос из базы. Для устранения речевых артефактов реализована простая нейронная сеть – перцептрон, которая с помощью функции engine.setProperty() реализует изменение свойств речи. Функция Speech() создает аудио-файл с готовой речью.

# **4. ТЕСТИРОВАНИЕ**

Для данного раздела были разработаны автономные тесты, ниже можно увидеть выдаваемые результаты модулей.

В первом тесте введен простой короткий текст.

Во втором тесте введен более длинный сложный текст.

В третьем тесте представлен вариант, когда введен некорректный текст.

## **4.1. Основной модуль генерации речи**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Входные данные** | **Ожидаемые  выходные данные** | **Реальные  выходные данные** | **Совпадение (Да/Нет)** |
| 1 | Привет, я рад что могу говорить | ['Привет, я рад что могу говорить'  ] | ['Привет, я рад что могу говорить'  ] | Да |
| 2 | В 1779 г. Петербургской Академией наук была объявлена ежегодная премия за создание устройства, способного воспроизводить пять русских гласных звуков. Лучшим образом эту задачу выполнил немецкий ученый Христиан Готлиб Кратценштейн | [‘В 1779 г. Петербургской Академией наук была объявлена ежегодная премия за создание устройства, способного воспроизводить пять русских гласных звуков. Лучшим образом эту задачу выполнил немецкий ученый Христиан Готлиб Кратценштейн’] | [‘В 1779 г. Петербургской Академией наук была объявлена ежегодная премия за создание устройства, способного воспроизводить пять русских гласных звуков. Лучшим образом эту задачу выполнил немецкий ученый Христиан Готлиб Кратценштейн’] | Да |
| 3 | палоыпршд аооа рпра у шао | [‘палоыпршд аооа рпра у шао’] | [‘палоыпршд аооа рпра у шао’] | Да |

### **4.2. Анализ текста**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Входные данные** | **Ожидаемые  выходные данные** | **Реальные  выходные данные** | **Совпадение (Да/Нет)** |
| 1 | ['Привет, я рад что могу говорить'  ] | ['Привет,', 'я', 'рад', 'что', 'могу', 'говорить'] | ['Привет,', 'я', 'рад', 'что', 'могу', 'говорить'] | Да |
| 2 | [‘В 1779 г. Петербургской Академией наук была объявлена ежегодная премия за создание устройства, способного воспроизводить пять русских гласных звуков. Лучшим образом эту задачу выполнил немецкий ученый Христиан Готлиб Кратценштейн’] | ['В', '1779', 'г.', 'Петербургской', 'Академией', 'наук', 'была', 'объявлена', 'ежегодная', 'премия', 'за', 'создание', 'устройства,', 'способного', 'воспроизводить', 'пять', 'русских', 'гласных', 'звуков.', 'Лучшим', 'образом', 'эту', 'задачу', 'выполнил', 'немецкий', 'ученый', 'Христиан', 'Готлиб', 'Кратценштейн'] | ['В', '1779', 'г.', 'Петербургской', 'Академией', 'наук', 'была', 'объявлена', 'ежегодная', 'премия', 'за', 'создание', 'устройства,', 'способного', 'воспроизводить', 'пять', 'русских', 'гласных', 'звуков.', 'Лучшим', 'образом', 'эту', 'задачу', 'выполнил', 'немецкий', 'ученый', 'Христиан', 'Готлиб', 'Кратценштейн'] | Да |
| 3 | [‘палоыпршд аооа рпра у шао’] | ['палоыпршд', 'аооа', 'рпра', 'у'] | ['палоыпршд', 'аооа', 'рпра', 'у'] | Да |

### **4.3. Выбор звуковых секций**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Входные данные** | **Ожидаемые  выходные данные** | **Реальные  выходные данные** | **Совпадение (Да/Нет)** |
| 1 | ['Привет,', 'я', 'рад', 'что', 'могу', 'говорить'] | Речь, читающая слова из списка | Речь, читающая слова из списка (можно прослушать по ссылке, представленной в разделе «Итоговая работа программы») | Да |
| 2 | ['В', '1779', 'г.', 'Петербургской', 'Академией', 'наук', 'была', 'объявлена', 'ежегодная', 'премия', 'за', 'создание', 'устройства,', 'способного', 'воспроизводить', 'пять', 'русских', 'гласных', 'звуков.', 'Лучшим', 'образом', 'эту', 'задачу', 'выполнил', 'немецкий', 'ученый', 'Христиан', 'Готлиб', 'Кратценштейн'] | Речь, читающая слова из списка | Речь, читающая слова из списка (можно прослушать по ссылке, представленной в разделе «Итоговая работа программы») | Да |
| 3 | ['палоыпршд', 'аооа', 'рпра', 'у'] | Невнятная речь, читающая слова из списка | Невнятная речь, читающая слова из списка  (можно прослушать по ссылке, представленной в разделе «Итоговая работа программы») | Да |

### **4.4. Генерация речи**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Входные данные** | **Ожидаемые  выходные данные** | **Реальные  выходные данные** | **Совпадение (Да/Нет)** |
| 1 | ['Привет,', 'я', 'рад', 'что', 'могу', 'говорить'] | Более качественная речь, читающая цельное предложение | Более качественная речь, читающая цельное предложение (можно прослушать по ссылке, представленной в разделе «Итоговая работа программы») | Да |
| 2 | ['В', '1779', 'г.', 'Петербургской', 'Академией', 'наук', 'была', 'объявлена', 'ежегодная', 'премия', 'за', 'создание', 'устройства,', 'способного', 'воспроизводить', 'пять', 'русских', 'гласных', 'звуков.', 'Лучшим', 'образом', 'эту', 'задачу', 'выполнил', 'немецкий', 'ученый', 'Христиан', 'Готлиб', 'Кратценштейн'] | Более качественная речь, читающая цельное предложение | Более качественная речь, читающая цельное предложение (можно прослушать по ссылке, представленной в разделе «Итоговая работа программы») | Да |
| 3 | [‘палоыпршд аооа рпра у шао’] | Невнятная речь, читающая слова из списка | Невнятная речь, читающая слова из списка (можно прослушать по ссылке, представленной в разделе «Итоговая работа программы») | Да |

# **5. РАБОТА ИТОГОВОГО ПРИЛОЖЕНИЯ**

В данном разделе представлена работа итогового приложения, ниже можно увидеть какими сообщениями сопровождается работа программы.

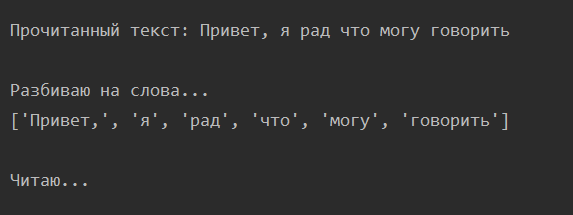


Рис. 3. Работа итогового приложения

ЗАКЛЮЧЕНИЕ

Данная технология является очень многогранной и стремительно развивается вместе с технологическим прогрессом. В будущем она будет так же востребована, как и технология искусственного интеллекта. С помощью данной работы была собрана основная информация для ознакомления с синтезом речи и начала работы с этой технологией.

В ходе работы были исследованы различные способы реализации синтезатора речи, включая процессы анализа текста, преобразования его в список слов и обучение нейронной сети. В результате исследований, были выбраны наиболее подходящие методы реализации и выполнено проектирование модульной структуры синтезатора речи. На основе модульной структуры была написана программа на языке программирована Python, реализующая простой синтезатор речи. Также было разработано и успешно пройдено тестирование.

Исходные тексты разработанных модулей и результаты тестирования доступны по ссылке <https://github.com/puffsi/NIR.git>.

Список литературы:

1. Соломенник А.И. Технология синтеза речи: история и методология исследований / Вестник Московского университета. Сер. 9, Филология, 2013, №6, С. 149-163
2. Калиев А., Рыбин С.В. Синтез речи: прошлое и настоящее / Компьютерные инструменты в образовании, 2019, №1, С. 5-28
3. Чистиков П.Г., Рыбин С.В. Проблемы естественности речевого сигнала в системах синтеза / Компьютерные инструменты в образовании, 2011, №1, С. 22-30
4. Рыбин С.В. СИНТЕЗ РЕЧИ Учебное пособие по дисциплине "Синтез речи", СПб: Университет ИТМО, 2014, 92 с.
5. Пугач А.С. Базовые принципы построения системы синтеза речи / Молодой ученый, 2017, №13 (147), С. 16-19
6. Людовик Т.В. Автоматический синтез нейтральной и выразительной речи / Штучний інтелект, 2010, № 1, С. 93-102
7. Кнеллер Э.Г. Анализ параметров речевого сигнала создающих восприятие элементарных звуков речи / Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея, – М.: Изд-во РГГУ, 2006, С. 220-223
8. Котов А.А. Модель эмоционального речевого поведения виртуального агента ролевой компьютерной игры / Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2006, С. 285-289
9. Лобанов Б.М., Цирульник Л.И., Пьорковска Б., Рафалко Я., Шпилевский Э. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках / Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2006, С. 357-363
10. Сечуйски, М., Острогонац, С., Сузич, С., Пекар, Д. Обучение просодической модели по данным в нейросетевом синтезе речи // *Труды СПИИРАН*, №*4*(59), C.193-215
11. Joo Y.-S., Bae H., Y. Kim Y.-I., Cho H.-Y. and Kang H.-G., Effective Emotion Transplantation in an End-to-End Text-to-Speech System, in IEEE Access, vol. 8, pp. 161713-161719, 2020, **DOI:**[10.1109/ACCESS.2020.3021758](https://doi.org/10.1109/ACCESS.2020.3021758)
12. Isewon I., Oyelade J., Oladipupo O., Design and Implementation of Text To Speech Conversion for Visually Impaired People in International Journal of Applied Information Systems (IJAIS), vol. 7, pp. 25-30, 2014, DOI: [10.5120/ijais14-451143](https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.5120%2Fijais14-451143)
13. Aida–Zade K.R., Ardil C. and Sharifova A.M., The Main Principles of Text-to-Speech Synthesis System in World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering, vol. 7, No:3, pp. 395-401, 2013
14. Pavlova E., Pavlov Y., Sproat R., Shih C., Jan P. H. van Santen, BELL LABORATORIES RUSSIAN TEXT-TO-SPEECH SYSTEM, at EUROSPEECH ’97 5th European Conference on Speech Communication and Technology Rhodes, Greece, September 22-25, 1997
15. Sefara T.J., Manamela M., The Development of Local Synthetic Voices for an Automatic Pronunciation Assistant, at Southern Africa Telecommunication Networks and Applications, George, September, 2016
16. Kenmochi H., Ohshita H., VOCALOID - commercial singing synthesizer based on sample concatenation, In *INTERSPEECH-2007*, 4011-4010
17. Aaron van den Oord, Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., Kalchbrenner N., Senior A., Kavukcuoglu K., WaveNet: A Generative Model for Raw Audio, Google DeepMind, London, UK, 19 September 2016
18. Ling Z.-H., Kang S.-Y., Zen H., Senior A., Schuster M., Qian X.-J., Meng H., Deng L. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends // IEEE Signal Processing Magazine. 2015. Vol. 32. № 3. P. 35–52. doi: 10.1109/MSP.2014.2359987

Приложение 1

В файле main\_speech\_generation.py реализуется чтения текста из файла Text.txt. Данный модуль вызывает функции импортированных файлов: text\_analys.py , sound\_choose.py , и speech\_generation.py. Выходным параметром является строка my\_string.

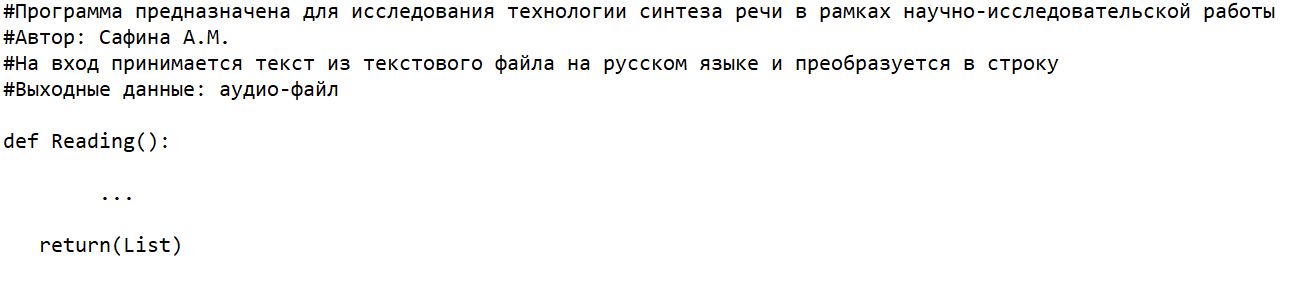


Рис. 4. Интерфейс Главного модуля генерации речи

**Приложение 2**

Файл text\_analys.py с помощью тривиального алгоритма реализуется разбиение строки my\_string на слова и создание списка этих слов. Входными данными является строка my\_string, полученная в модуле main\_speech\_generation.py. Выходными данными является список words.

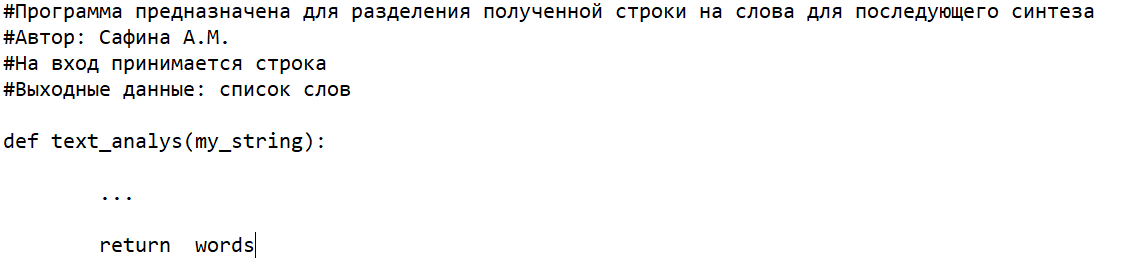
****

Рис. 5. Интерфейс модуля анализа текста

**Приложение 3**

Файл sound\_choose.py выполняет функцию озвучивания полученного списка слов в файле text\_analys.py. Для реализации этой функции используется модуль pyttsx 3 и голосовая база RH Voice. Входными данными является строка my\_string\_1, содержащая отдельные слова. Выходными данными является 0, так как результатом выполнения является воспроизведение аудио.

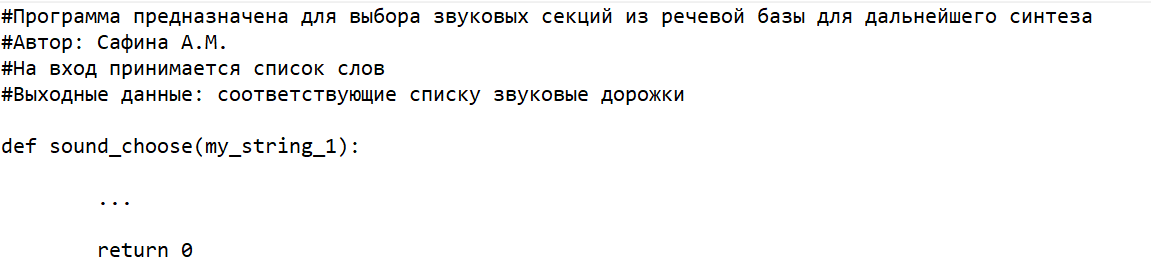
****

Рис. 6. Интерфейс модуля выбора звуковых секций

**Приложение 4**

Файл speech.py выполняет функцию озвучивания полученного списка слов в файле text\_analys.py и создания аудио-файла. Полученная в результате речь проходит через простую нейронную сеть – перцептрон, который регулирует параметры голоса для лучшего качества речи. Для реализации этой функции используется модуль pyttsx 3, голосовая база RH Voice, а так же простой алгоритм создания перцептрона. Входными данными является строка my\_string\_1, содержащая отдельные слова. Выходными данными является аудио-файл.

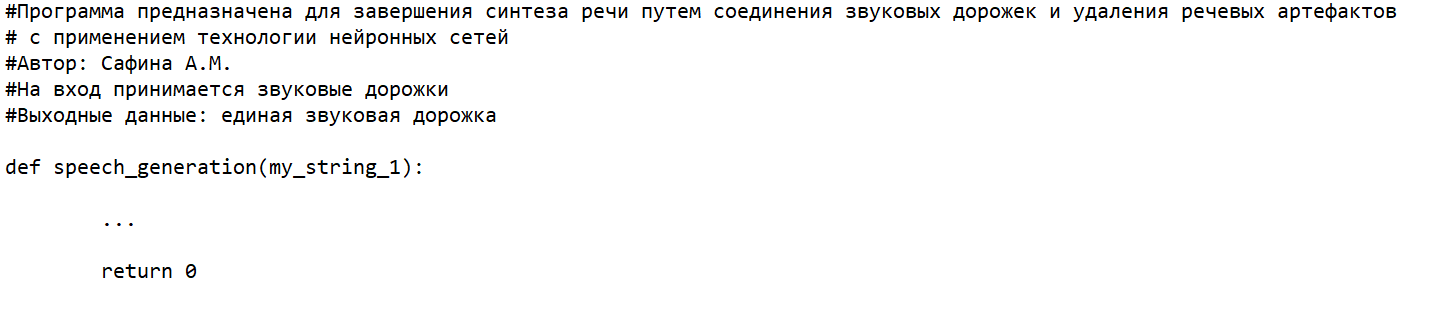


Рис. 7. Интерфейс модуля генерации речи