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

Студент: Овчинникова Анастасия Павловна

Научный руководитель:

доцент кафедры ИУ-7, к. ф.-м. н. Романова Татьяна Николаевна

# Актуальность

- Успешное развитие систем разговорного языка 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.
- Голосовые роботы берут на себя большую часть рутинных задач, например, в колл-центрах.
- Использование систем синтеза речи может повысить качество жизни слабослышащих, глухих, немых людей.

# Цель и задачи работы

**Цель**: разработка алгоритма для синтеза речи по произвольному тексту на русском языке и создание ПО для озвучивания текста на основе этого алгоритма.

**Задачи, которые следует решить для достижения поставленной цели**:

- провести аналитическое исследование существующих подходов к синтезу речи и выбрать наиболее подходящий;
- на основе проведенного исследования модифицировать выбранный метод для достижения поставленной цели;
- разработать алгоритм по модифицированному методу;
- создать программное обеспечение, реализующее данный алгоритм;
- разработать тесты для разработанного ПО;
- провести исследование разработанного ПО для оценки качества синтезируемой речи.

# Методы синтеза речи

| Название метода | Достоинства | Недостатки |
|---|---|---|
| **Параметрический** | • Очень высокое качество речи. | • Набор текстовых сообщений, которые необходимо озвучить, должен быть заранее известен. |
| **Компилятивный** | • Простота.<br>• Качество синтезируемой речи зависит только качества записи элементов синтеза. | • Большой объем памяти для хранения предварительно записанного словаря.<br>• Содержание и вариативность синтезируемой речи ограничена этим словарем. |
| **По фонетическим правилам** | • Позволяет синтезировать речь по заранее неизвестному тексту. | • Самый сложный из всех трех методов.<br>• Синтезированная речь имеет худшее качество, чем в предыдущих двух методах. |

Единственный метод, позволяющий озвучить произвольный заранее неизвестный текст, – синтез **по фонетическим правилам**.

# Исходные элементы синтеза

| Исходные элементы | Достоинства | Недостатки |
|---|---|---|
| **Микросегменты** | • Наиболее «детально» из всех вариантов учитывается контекст. | • Большой объем БД.<br>• Трудно выделить границу микросегментов.<br>• Много «склеек». |
| **Аллофоны** | • При использовании мини-набора объем БД будет небольшим. | • Трудно выделить границу аллофонов.<br>• Для улучшения качества синтеза необходимо сильно увеличить объем БД. |
| **Дифоны** | • Контекст учитывается более детально. | • Трудно выделить границу дифонов.<br>• Для улучшения качества синтеза необходимо сильно увеличить объем БД. |
| **Полуслоги** | • Контекст учитывается более детально. | • Трудно выделить границу полуслогов. |
| **Слоги** | • Меньше проблем с выделением границ. | • Большой объем БД. |

**Аллофоны** являются наилучшим компромиссом между возможным качеством синтезируемой речи и объемом БД исходных элементов.

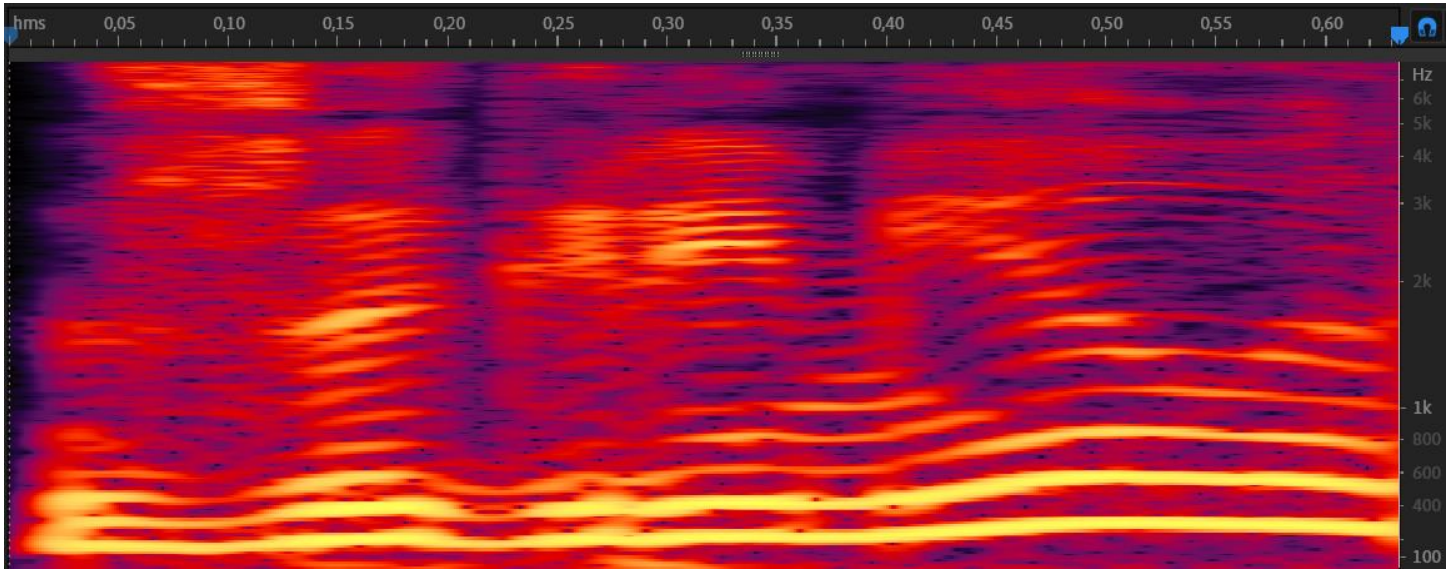
# Создание акустической базы данных

- Всего в русском языке насчитывается 42 фонемы.
- Обычно выделяют и используют два практически обоснованных варианта набора аллофонов: мини-набор (420 аллофонов) и макси-набор (4140 аллофонов).
- В данной работе используется мини-набор (420 аллофонов).
- Для создания акустической базы данных был использован корпус RUSLAN.

# Модификация метода аллофонного синтеза

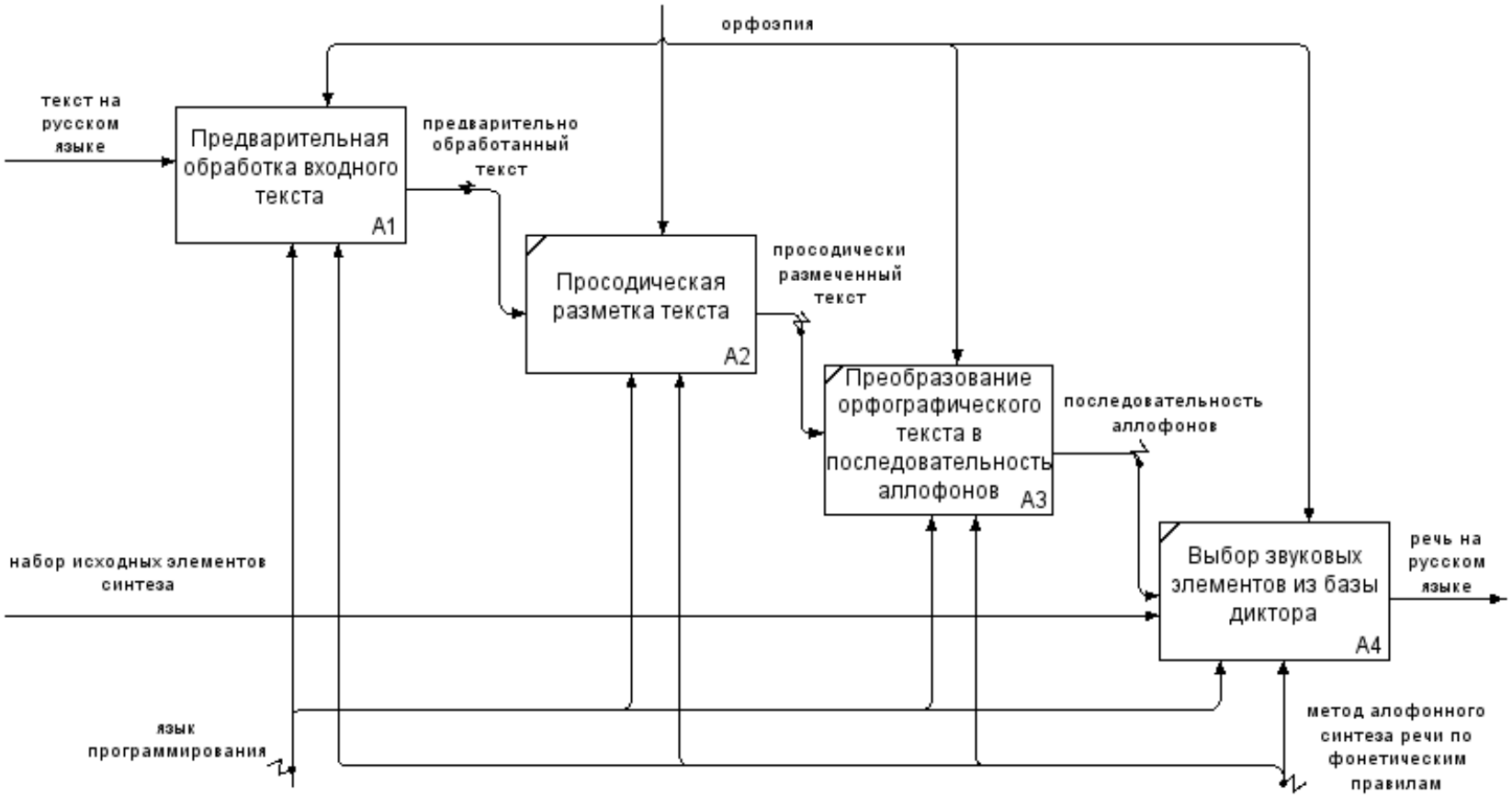
В потоке речи взаимовлияние соседних аллофонов друг на друга оказывается настолько сильным, что при создании БД аллофонов не всегда удается провести между ними четкую границу.

В данной работе в качестве исходных элементов синтеза использовать не только аллофоны, но  и более протяжённые фонетические сегменты (аллослоги).




Спектрограмма слова «загремела» (получена с помощью Adobe Audition).

# Функциональная модель реализации метода



# Лингвистический текстовый процессор


Схема алгоритма работы лингвистического текстового процессора

- Задачей лингвистического текстового процессора является предварительная обработка текста.
- Под дешифровкой аббревиатур здесь понимается установление правил их чтения, а не расшифровка слов, сокращением которых она является.

# Ограничения разработанного лингвистического текстового процессора (ЛТП)

Блок очистки текста удаляет из входного текста выделенные ограничения.

**Что ЛТП может делать:**

- обрабатывать собственно текст на русском языке, состоящий из предложений или отдельных слов и знаков препинания;
- дешифровать аббревиатуры;
- заменять «е» на «ё» в словах, где написана буква «е» вместо «ё».

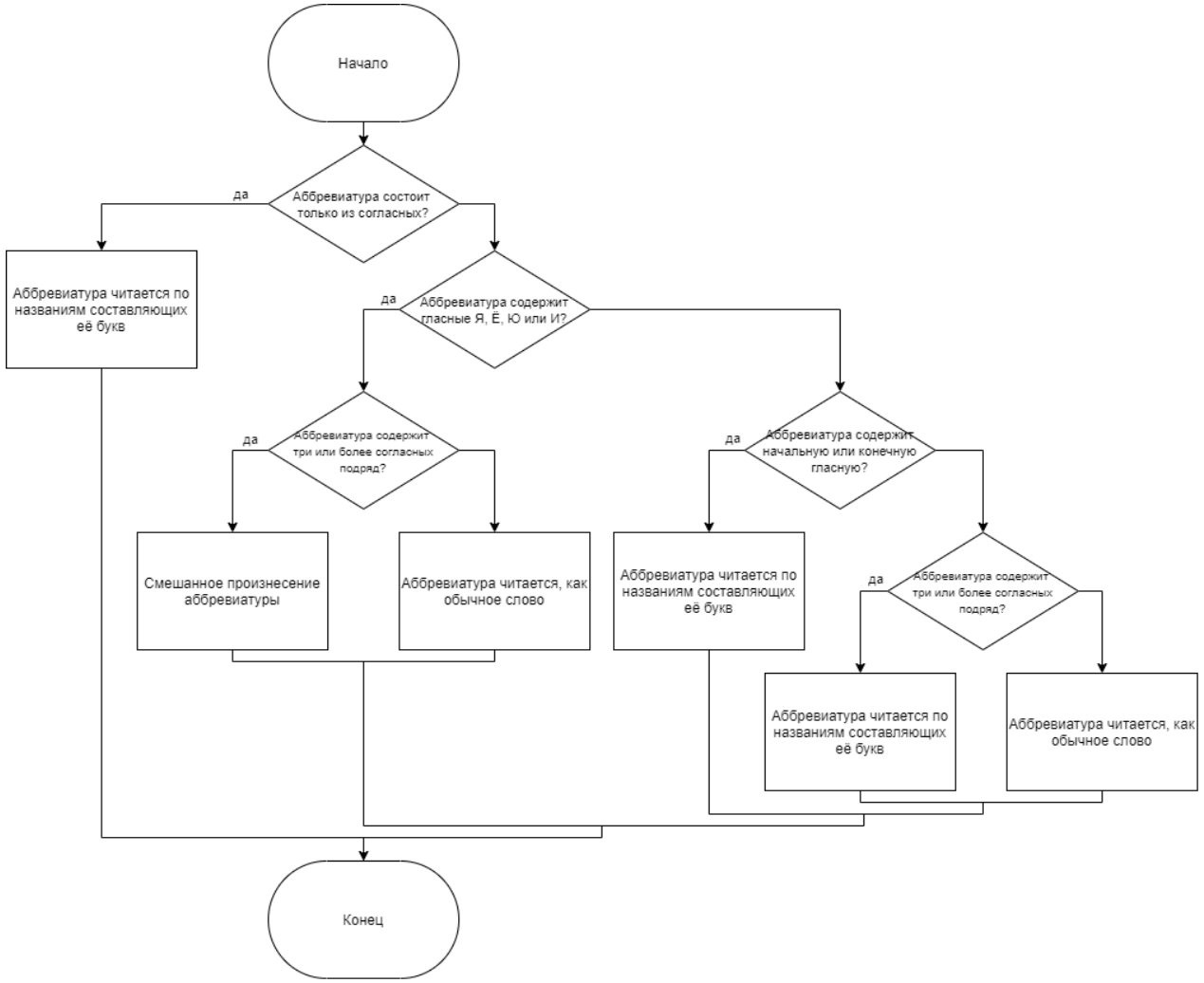
**Что ЛТП не может обрабатывать:**

- сокращения.

**Что ЛТП удаляет из текста:**

- иностранные слова;
- специальные символы;
- многоразрядные и дробные числа;
- телефонные номера;
- обозначения времени и даты;
- интернет-адреса;
- какую-либо разметку текста (например, просодическую);
- математические выражения.

# Схема разработанного алгоритма для дешифровки неизвестных аббревиатур



# Просодический процессор

- 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.
- Считается, что пунктуационные синтагмы ограничены следующими знаками препинания: [ ; ], [ : ], [ , ], [ - ], [ ( ], [ ) ], [ « ], [ « ], [ ,- ]. Если знак препинания стоит после сочинительного союза (и, да, но и, так и, а и др.), то граница синтагмы в этом месте не проводится.

# Фонетический процессор

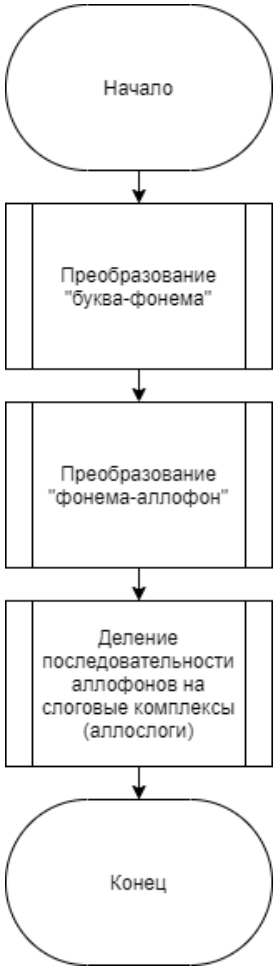



Схема алгоритма работы фонетического процессора

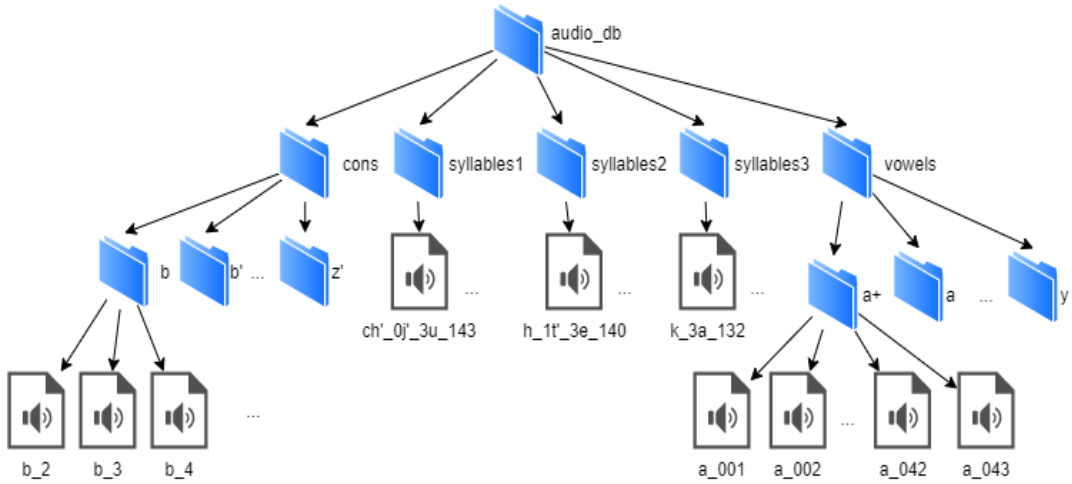
- Задачей фонетического процессора является преобразование орфографического текста в последовательность аллофонов и деление последовательности аллофонов на аллослоги.
- Фонетический процессор использует правила преобразования орфографического текста в фонемную последовательность (с учетом целого ряд сложившихся исключений).
- Аллофон является конкретной реализацией фонемы в речи. Аллофоны делятся на комбинаторные (определяются ближайшим контекстом фонему) и позиционные (определяются положением фонемы по отношению к ударному слогу в слове).

# Акустический процессор

- Задачей акустического процессора является выбор необходимых элементов из базы данных исходных элементов и их склейка в общий выходной аудиофайл.
- После склейки всех аллофонов на синтагму накладывается эффект fade in (постепенное увеличение громкости аудио-дорожки к ее концу) с помощью библиотеки pydub.
- Для того, чтобы речь звучала более равномерно, используется алгоритм интерполяции звуковых сигналов (увеличения частоты дискретизации сигнала в N раз) из библиотеки pydub.

# Организация хранения данных

- Для разработанной системы синтеза речи необходимо четыре словаря и БД исходных элементов синтеза.
- Для хранения словарей используется NoSQL база данных типа «ключ-значение».
- Исходные элементы хранятся в файловой системе в виде звуковых файлов.




Структура хранилища исходных элементов

# Результаты тестирования

Для тестирования разработанного программного обеспечения были разработаны unit-тесты. Все тесты были пройдены.

| № | Входные данные | Ожидаемый результат |
|---|---|---|
| 1 | "ПРИНЁ+С_ИГРУ+ШКУ" | ["p", "r'", "i", "n'", "o", "+", "s", "_", "y", "g", "r", "u", "+", "sh", "k", "u"] |
| 2 | "ОТЦА+" | ["a", "c", "c", "a", "+"] |

Примеры входных данных и ожидаемых результатов тестов для блока преобразования буква-фонема.

| № | Входные данные | Ожидаемый результат |
|---|---|---|
| 1 | ["#", "u"] | ["pause_1", "u_100"] |
| 2 | ["_", "a", "_"] | ["pause_0", "a_100", "pause_0"] |

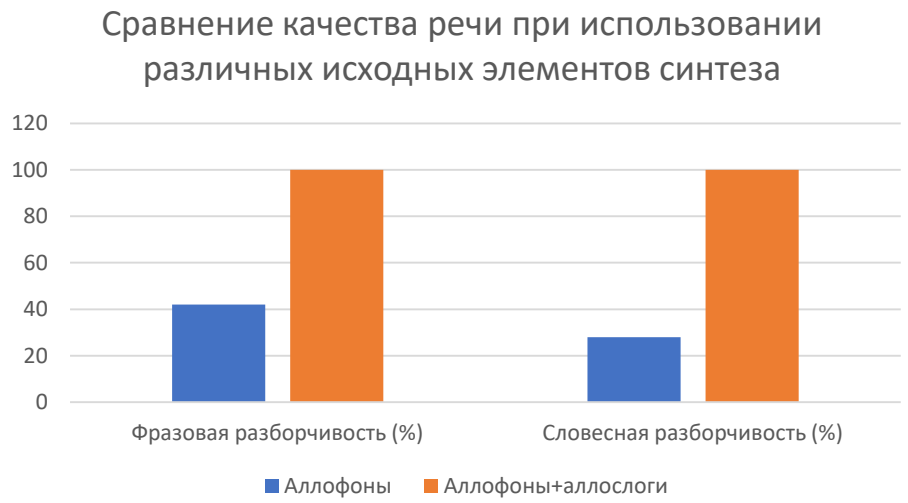
Примеры входных данных и ожидаемых результатов тестов для блока преобразования фонема-аллофон.

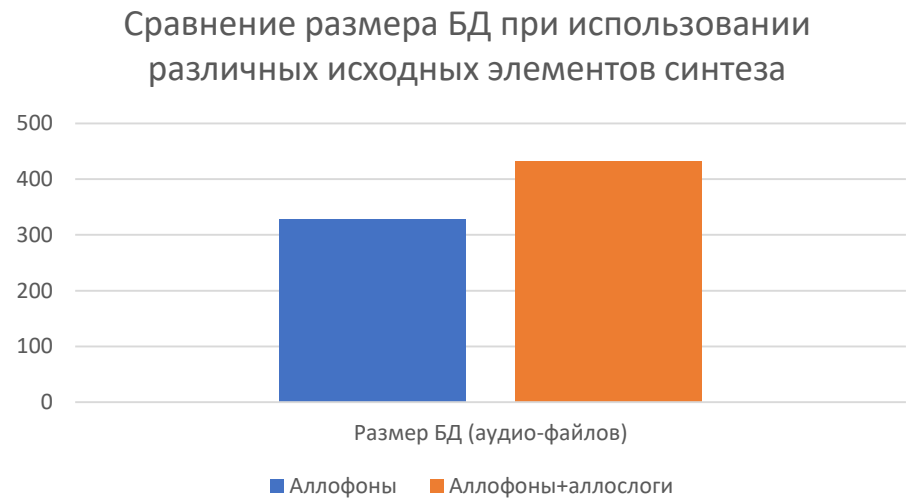
| № | Входные данные | Ожидаемый результат |
|---|---|---|
| 1 | "КО+ЛОСА" | [["k", "o"], ["l", "a"], ["s", "a"]] |
| 2 | "А+ММО" | [["a"], ["m", "m", "a"]] |

Примеры входных данных и ожидаемых результатов тестов для блока деления на открытые слоги.

# Сравнение исходного и модифицированного методов

| Исходные элементы синтеза | Фразовая разборчивость | Смысловая разборчивость | Словесная разборчивость | Естественность (0-5 баллов) |
|---|---|---|---|---|
| Аллофоны | 42% | Неудовлетворительная | 28% | < 1,7 балла |
| Аллофоны+аллослоги | 100% | Отличная | 100% | 3,0 – 3,4 балла |



# Заключение

В результате выполненной работы были решены все поставленные задачи.

Было создано ПО, которое:

- позволяет синтезировать речь по произвольному входному тексту, качество речи достаточно высоко: речь понятна, но не эмоциональна (нет интонации);
- позволяет повысить коммуникативные возможности слабослышащих, глухих, немых людей;
- может быть успешно использовано в различных отраслях при создании голосовых роботов.

# Дальнейшее развитие

- Улучшение качества синтезируемой речи за счет увеличения объема БД исходных элементов.
- Усовершенствование алгоритмов обработки входного текста.
- Поддержка пользовательской разметки (расстановка словесного и синтагматического ударения).
- Интеграция синтаксического и семантического анализатора с целью уменьшения ограничений, накладывающихся на исходный текст.
- Добавление интонационного оформления речи.

Спасибо за внимание!

# Критерии оценки качества синтезируемой речи

- **Словесная разборчивость** – оценивается количество правильно воспринятых никак не связанных между собой слов;
- **Фразовая разборчивость** – оценивается количество правильно распознанных фраз;
- **Смысловая разборчивость** – оценивается понимание содержания речи.
- Фразовая разборчивость должна быть на уровне 98-100%. При оценке словесной разборчивости требуется результат не менее 99%.

# Градации качества при оценке смысловой разборчивости речи

- отлично – полное понимание речи, отсутствие переспросов;
- хорошо – понимание речи полное, но возможны переспросы необычных слов, фамилий и терминов;
- удовлетворительно – переспросы отдельных слов и фраз, но в целом восприятие речевой информации правильное;
- неудовлетворительно – отдельные слова непонятны даже при переспросе;
- плохо – смысл речевой информации понимается с трудом.

# Градации качества при оценке естественности речи

- 4,6 – 5,0 баллов – естественность звучания речи, полное отсутствие искажений;
- 4,0 – 4,5 – естественность звучания речи, малозаметные искажения;
- 3,5 – 3,9 – естественность звучания речи, слабое постоянное присутствие искажений;
- 3,0 – 3,4 – незначительное нарушение естественности, заметное присутствие искажений;
- 2,5 – 2,9 – заметное нарушение естественности, присутствие искажений или помех;
- 1,7 – 2,4 – существенное искажение естественности, постоянное присутствие искажений или помех;
- < 1,7 – сильные искажения, механический голос, потеря естественности.

# Контрольные примеры

- «Темной зимней ночью несколько английских судов направлялись к бухте».
- «Загремела с визгом якорная цепь».
- «Молния блистала все чаще и ярче».
- «На пароходе зажгли электричество».
- «Они пошли в гостиную к роялю».
- «Жара понемногу спадала».
- «Жарко от летнего солнца и от теплой земли».

- Шаромыжник.
- Свентицкой.
- Смоковников.
- Геннисон.
- Энниок.
- Эбергайль.
- Коломб.

# Основные определения

- **Фонема** – абстрактная единица языка, служащая для формирования означающего языковых знаков. Фонемы участвуют в различении звуковых оболочек. Фонема является абстрактной единицей языка, а ее конкретной реализацией в речи является звук.
- **Аллофон** — реализация фонемы, один из ее вариантов, обусловленный конкретным фонетическим окружением. В отличие от фонемы, является не абстрактным понятием, а конкретным речевым звуком.
- **Дифон** — сегмент речи между серединами соседних аллофонов.
- **Полуслог** — сегмент, содержащие половину согласного и половину примыкающего к нему гласного.
- **Аллослог** – слоговой сегмент с учетом позиционной и комбинаторной аллофонии.

# Основные определения

- **Фонетическое слово** – отрезок речевой цепи, объединяемый одним (словесным) ударением.
- **Синтагма** – последовательность слов, связанных синтаксической связью, которые представляют из себя цельные просодические единицы. На границах синтагмы происходит смена просодического оформления.
- **Просодические параметры** подразделяются на тональные (мелодика, изменения частоты основного тона), количественно-динамические (паузы, длительность, темп и интенсивность), артикуляционные (раствор рта, назализация, смещение язычной артикуляции) и фонационные (различные типы голоса).

# Диаграмма классов