Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»

Кафедра интеллектуальных информационных технологий

**ОТЧЕТ**

**по производственной практике**

**тема практики “Разработка синтезатора речи”**

|  |  |
| --- | --- |
| Выполнил студент группы 321702: | Пинчуков А. Р. |
| Руководитель от производства: | Н. |
| Руководитель от кафедры: | Бойко И. М. |
|  |  |

Минск 2018

**СОДЕРЖАНИЕ**

[**ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ** 3](#_Toc515099991)

[**ВВЕДЕНИЕ** 4](#_Toc515099992)

[**1.** **АНАЛИЗ, СУЩЕСТВУЮЩИЙ ПОДХОДОВ СИНТЕЗА РЕЧИ** 5](#_Toc515099993)

[**1.1 Конкатенативный синтез** 5](#_Toc515099994)

[**1.1.1** **Синтез выбора единиц** 5](#_Toc515099995)

[**1.1.2** **Синтез дифонов.** 5](#_Toc515099996)

[**1.1.3** **Синтез доменов** 6](#_Toc515099997)

[**1.2 Формантный синтез** 6](#_Toc515099998)

[**1.2.1 Артикуляционный** **синтез** 7](#_Toc515099999)

[**1.2.2 Синтез на основе HMM** 7](#_Toc515100000)

[**1.2.3 Синтетический синтез** 7](#_Toc515100001)

[**2.** **РАЗРАБОТКА преобразователя речи в текст** 8](#_Toc515100002)

# **ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ**

В отчете используются следующие условные обозначения:

ПО – программное обеспечение

БД – база данных

TTS – text to speech

**ВВЕДЕНИЕ**

Синтез речи - искусственное воссоздание человеческой речи. Компьютерная система, используемая для этой цели, называется речевым компьютером или речевым синтезатором и может быть реализована в программных или аппаратных продуктах. Система преобразования текста в речь (TTS) преобразует текст обычного языка в речь; другие системы предоставляют символические лингвистические представления, такие как фонетические транскрипции в речи. [1]

Синтезированная речь может быть создана путем объединения фрагментов записанной речи, которые хранятся в базе данных. Системы отличаются размером хранимых речевых единиц; система, в которой хранятся телефоны или дифоны, обеспечивает наибольший выходной диапазон, но может не иметь четкости. Для конкретных доменов использования хранение целых слов или предложений позволяет получить высококачественную продукцию. Альтернативно, синтезатор может включать в себя модель голосового тракта и других характеристик голоса человека для создания полностью «синтетического» речевого выхода. [2]

Качество речевого синтезатора оценивается по его сходству с человеческим голосом и его способностью быть понятным. Интеллектуальная программа «текст в речь» позволяет людям с нарушениями зрения или чтениям слушать письменные слова на домашнем компьютере. Многие компьютерные операционные системы включали речевые синтезаторы с начала 1990-х годов.

Система преобразования текста в речь (или «двигатель») состоит из двух частей: [3] интерфейсной и внутренней. Интерфейс имеет две основные задачи. Во-первых, он преобразует исходный текст, содержащий символы, такие как числа и аббревиатуры, в эквивалент выписанных слов. Этот процесс часто называют нормализацией текста, предварительной обработкой или токенизацией. Затем интерфейс назначает фонетические транскрипции каждого слова, делит и маркирует текст в просодические единицы, такие как фразы и предложения. Процесс присвоения фонетических транскрипций словам называется преобразованием текста в фонеме или графем-фонеме. Фонетические транскрипции и информация просодии (тон, интонация) вместе составляют символическое лингвистическое представление, которое выводится интерфейсом. Внутренняя часть, часто называемый синтезатором, затем преобразует символическое лингвистическое представление в звук. В некоторых системах эта часть включает в себя расчет целевой просодии (контур высоты тона, длительности фонем), [4], который затем накладывается на выходную речь.

1. **АНАЛИЗ, СУЩЕСТВУЮЩИЙ ПОДХОДОВ СИНТЕЗА РЕЧИ**

Важнейшими качествами системы синтеза речи являются естественность и разборчивость. [21] Естественность описывает, насколько близко результат звучит как человеческая речь, в то время как разборчивость - это легкость, с которой понимается вывод. Идеальный синтезатор речи является естественным и понятным. Системы синтеза речи обычно стараются максимизировать обе характеристики.

Двумя основными технологиями, генерирующими синтетические речевые сигналы, являются конкатенативный синтез и формантный синтез. Каждая технология имеет свои сильные и слабые стороны, и предполагаемое использование системы синтеза обычно определяет, какой подход используется.

**1.1 Конкатенативный синтез**

Конкатенативный синтез основан на конкатенации сегментов записанной речи. Существует три основных подтипа конкатенативного синтеза.

* + 1. **Синтез выбора единиц**

В синтезе выбора единиц используются большие базы данных записанной речи. Во время создания базы данных каждое записанное высказывание сегментируется в некоторые или все из следующих: отдельные фонемы, дифоны, полуфонемы, слоги, морфемы, слова, фразы и предложения. Как правило, деление на сегменты выполняется с использованием специально модифицированного распознавателя речи, установленного в режим принудительного выравнивания с некоторой ручной коррекцией, с использованием визуальных представлений, таких как форма волны и спектрограмма. [22] Затем создается индекс единиц в речевой базе данных на основе сегментированных и акустических параметров, таких как основная частота (шаг), продолжительность, положение в слоге и соседние фонемы. Во время выполнения желаемое целевое высказывание создается путем определения лучшей цепочки единиц-кандидатов из базы данных (выбор единицы). Этот процесс обычно достигается с использованием специально взвешенного дерева решений.

* + 1. **Синтез дифонов.**

Синтез дифонов использует минимальную речевую базу данных, содержащую все дифоны (переходы звука), встречающиеся на языке. В синтезе дифонов в речевой базе данных содержится только один пример каждого дифона. Во время выполнения целевая просодия предложения накладывается на эти минимальные единицы с помощью цифровых технологий обработки сигналов, таких как линейное предсказательное кодирование, PSOLA [26] или MBROLA. [27] или более поздние методы, такие как изменение высоты тона в исходной области с использованием дискретного косинусного преобразования [28]. Синтез Дифона страдает от звуковых сбоев конкатенативного синтеза и роботочувствительной природы формантного синтеза и имеет мало преимуществ любого другого подхода, кроме маленького размер БД. Таким образом, его использование в коммерческих приложениях снижается, хотя он по-прежнему используется в исследованиях, поскольку существует множество свободно доступных программных реализаций.

* + 1. **Синтез доменов**

Синтез, специфичный для домена, объединяет предварительно записанные слова и фразы для создания полных высказываний. Он используется в приложениях, где множество текстов, которые будет выводиться системой, ограничено определенным доменом, например, сообщениями о расписании маршрута или метеорологическими отчетами. [29] Технология очень проста в применении и долгое время используется в коммерческих целях в таких устройствах, как говорящие часы и калькуляторы. Уровень естественности этих систем может быть очень высоким, поскольку разнообразие типов предложений ограничено, и они тесно соответствуют просодии и интонации оригинальных записей.

Поскольку эти системы ограничены словами и фразами в своих базах данных, они не являются универсальными и могут синтезировать только комбинации слов и фраз, с которыми они были предварительно запрограммированы. Однако смешение слов в естественном разговоре может все же вызвать проблемы, если не учитывать множество вариаций.

**1.2 Формантный синтез**

Формантный синтез не использует человеческие образцы речи во время выполнения. Вместо этого синтезированный речевой вывод создается с использованием аддитивного синтеза и акустической модели (синтез физического моделирования) [30]. Параметры, такие как основные частоты, голоса и уровни шума, со временем изменяются для создания формы искусственной речи. Этот метод иногда называют синтезом на основе правил; однако многие конкатенативные системы также имеют компоненты на основе правил. Многие системы, основанные на технологии синтеза формантов, создают искусственную, роботизирующую речь, которая никогда не принимается за человеческую речь. Однако максимальная естественность не всегда является целью системы синтеза речи, а системы синтеза формантов имеют преимущества перед конкатенационными системами. Формант-синтезированная речь может быть понятной даже на очень высоких скоростях, избегая акустических сбоев, которые обычно поражают конкатенативные системы. Ускоренная синтезированная речь используется слабовидящими, чтобы быстро перемещаться по компьютерам с помощью устройства чтения с экрана. Формант-синтезаторы обычно представляют собой более простые программы, чем конкатенационные системы, потому что у них нет базы данных образцов речи. Поэтому их можно использовать во встроенных системах, где память и мощность микропроцессора особенно ограничены. Поскольку системы на основе формант имеют полный контроль над всеми аспектами выходной речи, можно выводить самые разнообразные прозы и интонации, передавая не только вопросы и заявления, но и множество эмоций и тонов голоса.

**1.2.1 Артикуляционный** **синтез**

Артикуляционный синтез относится к вычислительным методам синтеза речи на основе моделей человеческого голосового тракта и происходящих там артикуляционных процессов.

До недавнего времени модели синтеза артикуляции не были включены в коммерческие системы синтеза речи. Примечательным исключением является система на основе NeXT, первоначально разработанная и продаваемая Trillium Sound Research, побочной компанией Университета Калгари, где проводилась большая часть оригинальных исследований. Система, впервые выпущенная на рынок в 1994 году, обеспечивает полное преобразование текста в речь на основе артикулятора, используя волновод или аналоговый канал линии передачи данных орального и носового тракта человека, контролируемый «отличительной моделью региона» Карре.

Более поздние синтезаторы, разработанные Хорхе С. Лусеро и его коллегами, включают в себя модели биомеханики вокальной складки, гортанной аэродинамики и распространения акустических волн в бронхи, траке, назальных и ротовых полостях и, таким образом, представляют собой полные системы моделирования речи на основе физики. [ 34] [35]

**1.2.2 Синтез на основе HMM**

Синтез на основе HMM представляет собой метод синтеза, основанный на скрытых марковских моделях, также называемый статистическим параметрическим синтезом. В этой системе HMM моделируются частотным спектром (голосовым трактом), основной частотой (источником голоса) и продолжительностью (просодией) речи. Речевые сигналы генерируются самими HMM на основе критерия максимального правдоподобия [36].

**1.2.3 Синтетический синтез**

Синтетическийсинтез - это метод синтеза речи путем замены формант (основных полос энергии) чистыми тональными свистами [37].

Для разработки преобразователя текста в речь был выбран конкатенативный синтез, на основе синтеза выбора единиц. За единицу принимались фонемы.

1. **РАЗРАБОТКА ПРЕОБРАЗОВАТЕЛЯ ТЕКСТА В РЕЧЬ**

Для разработки преобразователя был выбран язык программирования Java, т.к. он легко переносим на различные платформы, и достаточно быстр в работе.

Для ввода текста для преобразования был разработан пользовательский интерфейс, позволяющий ввести текст и после нажатия на кнопку “Run” прослушать уже преобразованный текст. На рисунке 2.1 продемонстрирован пользовательский интерфейс.

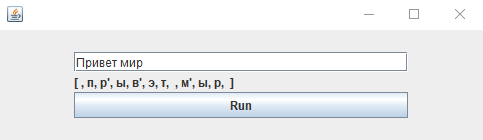


Рисунок 2.1 – Пользовательский интерфейс преобразователя текста в речь;

Алгоритм разбора текста на фонемы:

1.

На листенге 2.1 предоставлен код, разбивающий тек на фонемы.

public static List<String> getPhonemes(String text) {  
 List<String> phonemes = new ArrayList<>();  
 String lowerString = " " + text.toLowerCase() + " ";  
  
 for (int i = 0; i < lowerString.length(); i++) {  
  
 String letter = lowerString.substring(i, i + 1);  
 if (letter.equals("ь") || letter.equals("ъ")) {  
 continue;  
 }  
 if (letter.equals(" ")) {  
 phonemes.add(letter);  
 continue;  
 }  
  
 String nextLitter = null;  
 String prevLitter = null;  
  
 if (i + 1 < lowerString.length()) {  
 nextLitter = lowerString.substring(i + 1, i + 2);  
 }  
 if (i - 1 >= 0) {  
 prevLitter = lowerString.substring(i - 1, i);  
 }  
  
 if (!*isVowels*(letter)) {  
 if (letter.equals("ж")) {  
 phonemes.add("ж");  
 } else if (letter.equals("ш")) {  
 phonemes.add("ш");  
 } else if (letter.equals("ц")) {  
 phonemes.add("ц");  
 } else if (letter.equals("ч")) {  
 phonemes.add("ч");  
 } else if (letter.equals("щ")) {  
 phonemes.add("щ");  
 } else {  
 if (nextLitter != null && *isSoftness*(nextLitter)) {  
 phonemes.add(letter + "'");  
 } else {  
 phonemes.add(letter);  
 }  
 }  
 continue;  
 }  
  
 if (!*isIotovayaConsonant*(letter)) {  
 if (prevLitter != null && !*isVowels*(prevLitter)) {  
 phonemes.add("ы");  
 } else {  
 phonemes.add(letter);  
 }  
 continue;  
 }  
  
 if (prevLitter != null && (*isVowels*(prevLitter) ||  
 (prevLitter).equals(" ") || (prevLitter).equals("ъ") || (prevLitter).equals("ь")  
 )) {  
 if (letter.equals("е")) {  
 phonemes.add("й");  
 phonemes.add("э");  
 } else if (letter.equals("ё")) {  
 phonemes.add("й");  
 phonemes.add("о");  
 } else if (letter.equals("ю")) {  
 phonemes.add("й");  
 phonemes.add("у");  
 } else if (letter.equals("я")) {  
 phonemes.add("й");  
 phonemes.add("а");  
 } else if (letter.equals("и")) {  
 phonemes.add("й");  
 phonemes.add("ы");  
 }  
  
 } else if (letter.equals("е")) {  
 phonemes.add("э");  
 } else if (letter.equals("ё")) {  
 phonemes.add("о");  
 } else if (letter.equals("ю")) {  
 phonemes.add("у");  
 } else if (letter.equals("я")) {  
 phonemes.add("а");  
 } else if (letter.equals("и")) {  
 phonemes.add("ы");  
 }  
 }  
  
 return phonemes;  
}

Листинг 2.1 – Код разбора текста на фонемы

1. **3 часит**

**СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ**

1. Allen, Jonathan; Hunnicutt, M. Sharon; Klatt, Dennis (1987). From Text to Speech: The MITalk system. Cambridge University Press. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*0-521-30641-8*](https://en.wikipedia.org/wiki/Special:BookSources/0-521-30641-8).
2. [**Jump up^**](https://en.wikipedia.org/wiki/Speech_synthesis#cite_ref-2) Rubin, P.; Baer, T.; Mermelstein, P. (1981). "An articulatory synthesizer for perceptual research". Journal of the Acoustical Society of America. **70** (2): 321–328. [*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1121/1.386780*](https://doi.org/10.1121%2F1.386780)
3. van Santen, Jan P. H.; Sproat, Richard W.; Olive, Joseph P.; Hirschberg, Julia (1997). Progress in Speech Synthesis. Springer. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*0-387-94701-9*](https://en.wikipedia.org/wiki/Special:BookSources/0-387-94701-9).
4. [**Jump up^**](https://en.wikipedia.org/wiki/Speech_synthesis#cite_ref-4) Van Santen, J. (April 1994). "Assignment of segmental duration in text-to-speech synthesis". Computer Speech & Language. **8** (2): 95–128. [*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1006/csla.1994.1005*](https://doi.org/10.1006%2Fcsla.1994.1005).
   1. Taylor, Paul (2009). *Text-to-speech synthesis*. Cambridge, UK: Cambridge University Press. p. 3. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [9780521899277](https://en.wikipedia.org/wiki/Special:BookSources/9780521899277).