



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»

---

## Научно-исследовательская работа

Тема Анализ метода для распознавания слитной речи у людей с дефектом речи

Студент Козлова И.В.

Группа ИУ7-52Б

Научный руководитель Кивва К.А.

Москва — 2021 г.

# Введение

Речь – это самый распространенный вид индивидуального общения. Обработка речи – это изучение языковых сигналов. Сигналы обрабатываются в цифровой версии, поэтому обработку речи можно рассматривать как уникальный случай цифровой обработки сигналов [1].

Автоматическое распознавание речи – это процесс, в котором из входного речевого сигнала извлекаются необходимые признаки, затем с помощью этих признаков определяются слова/фразы, которые поступили на вход, структурная схема системы автоматического распознавания речи представлена на рисунке 1. То есть такая система позволяет компьютеру понимать слова, которые произносит человек [2].

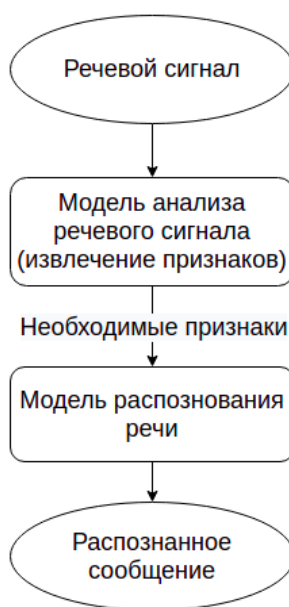


Рисунок 1 – Структурная схема системы автоматического распознавания речи

В современном мире существует множество технических средств, которые могут воспринимать произносимые речевые сообщения: мобильные телефоны, автомобили, компьютеры и др. Создание приложений, с помощью

которых машины могут разговаривать с человеком, особенно правильно реагируя на разговорную речь, давно начало интересоваться ученых и инженеров. Однако в настоящее время такая технология оптимизирована не для всех пользователей.

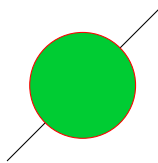
Системы автоматического распознавания речи широко применяются в медицинских исследованиях, например, когда требуется управлять автономными аппаратами. Важной областью применения систем автоматического распознавания речи является помощь людям с инвалидностью, как для людей с нарушениями речи, так и с проблемами опорно-двигательного аппарата [3].

**Цель данной научно-исследовательской работы** – это обоснованный выбор метода для распознавания слитной речи у людей с дефектом речи.

Для достижения поставленной цели необходимо решить следующие задачи:

- проанализировать классификацию систем автоматического распознавания речи;
- рассмотреть возможные дефекты и нарушения речи;
- проанализировать методы извлечения признаков (частотной характеристики);
- обосновать выбор метода извлечения признаков.

# 1 Аналитическая часть




./inc/img/2.pdf

Рисунок 1.1 – figure title

## 1.1 Классификация САРР

Системы автоматического распознавания речи (САРР, англ. ASR – Automatic Speech Recognition) помогают машинам интерпретировать устную речь и автоматизировать задачи человека, например поиск в интернете, набор текста и тд. Одним из наиболее сложных моментов в разработке таких систем является довольно широкая междисциплинарность задачи, то есть затрагиваются вопросы теории обработки сигналов, математического анализа, психологии, теории коммуникаций, а также лингвистики.

Системы автоматического распознавания речи можно классифицировать по основным аспектам [4]. К таким аспектам можно отнести следующие.

- Тип речи

- Спонтанная речь.

Спонтанную речь можно рассматривать как речь, которая звучит естественно (с эмоциями, с внезапными паузами).

- Изолированная речь.

Непрерывная речь – это обычная человеческая речь без пауз между словами. Этот вид значительно затрудняет машинное понимание речи.

- Связные слова.

Такие слова требуют минимальные паузы между высказываниями. Речь должна течь плавно.

- Изолированная речь.

Такие системы направлены на распознавания конкретных голосовых команд, получаемых от пользователя.

- Размер словаря.

- Маленькие словари.

САРР с маленькими словарями (чаще до 500-1000 слов) необходимы для распознавания команд, получаемых от пользователя.

- Большие словари.

Такие словари чаще всего используются в САРР слитной речи. Размеры достигают до десятков тысяч слов [5].

- Потребительские качества.

- Дикторозависимые.

К классу систем, зависящих от диктора относятся системы, которые требуют предварительного обучения и в его процессе настраиваются на определенного диктора. При смене диктора в таких системах возникает необходимость полной перенастройки.

- Дикторонезависимые.

К классу систем, независимых от диктора относятся системы, которые работают вне зависимости от того, кто выступает в качестве диктора. Данные системы имеют возможность распознавания речи любого диктора и не нуждаются в предварительном обучении.

- Структурные единицы.

В качестве структурных единиц могут выступать фразы, слова, фонемы. Системы, которые распознают речь, используя целые слова или фразы, называются системами распознавания речи по шаблону. Создание таких систем менее трудоемко, чем системы основанные на базе выделения лексических элементов (в таких системах структурными единицами являются фонемы).

- Принцип выделения структурных единиц.

В современных САРР используются несколько подходов для выделения структурных единиц из потока речи.

- Фурье-анализ (Жан-Батист Жозеф Фурье – французский математик и физик).

Данный анализ предполагает разложение исходной периодической функции в ряд, в результате чего исходная функция может быть представлена как суперпозиция синусоидальных волн различной частоты [6].

- Вейвлет-анализ (от англ. wavelet – ”маленькая волна”).

Данный анализ раскладывает исходный сигнал в базис функций, которые характеризуют как частоту, так и время [6].

- Кепстральный анализ.

Данный анализ основан на выделении кепстральных коэффициентов на мел-шкале, называемых мел-частотными кепстральными коэффициентами. Кепстр – это дискретно-косинусное преобразование амплитудного спектра сигнала в логарифмическом масштабе. Мел – единица высоты звука [7].

- Алгоритм распознавания

- Скрытые Марковские модели (СММ).

СММ – это модель, состоящая из  $N$  состояний, в каждом система может принимать одно из  $M$  значений какого-либо параметра. Матрицей  $A = a_{ij}$  задаются вероятности переходов между состояниями из  $i$  состояния в  $j$  состояние. Вектором  $B = b_j(k)$  задается вероятность выпадения какого-либо из  $M$  значений параметра в каждом из  $N$  состояний (выпадения  $k$  значения параметра в  $j$  состоянии). Вероятность того, что в начальный момент система окажется в  $i$  состоянии определяется вектором  $\pi = \pi_i$ . Таким образом, СММ называется тройка  $\lambda = A, B, \pi$ .

Модель называется ”скрытой”, потому что последовательность, в которой пребывала система неважна. Другими словами такая

система выступает в роли "черного ящика", на вход которого поступает последовательность параметров, а на выход ожидается модель, которая с максимальной вероятностью генерирует такую последовательность.

- Динамическое искажение времени.

Алгоритм динамического искажения времени (Dynamic Time Warping – DTW) является методикой эластичного сравнения вектора наблюдений с хранящимся шаблоном. По-другому можно сказать, что это мера подобия временных рядов, которая минимизирует эффекты временного сдвига, различного течения времени, а также обеспечивает непрерывное преобразование временных рядов для того, чтобы обнаружить одинаковые формы с различными фазами.

- Нейронные сети.

С помощью нейронных сетей можно создавать самообучаемые и обучаемые системы распознавания речи. Некоторые факторы, которым должны отвечать такие системы: возможность контроля своих действий с последующей коррекцией, разработка системы заключается только в построении архитектуры системы.

- Назначение

- Командные системы.

Такие системы используют распознавания по шаблону (фразе или слову).

- Системы диктовки.

Такие системы требуют более точного распознавания, то есть выделение лексических элементов. Также при интерпретации произнесенной фразы система полагается не только на то, что произносилось в данный момент, но и на фразы, сказанные ранее.

## 1.2 Сложности в работе САРР

Распознавание речи - это задача, усложненная тем, что речь человека характеризуется высокой степенью изменчивости [8].

Причины этого следующие:

- для одного и того же диктора произношения одних и тех же звуков (слов, фраз) будут отличаться длительностью произношения, интонацией. Часто это связано с изменением физического или эмоционального состояния человека, его настроения или условий, в которых он находится;
- произношение фонем сильно зависит от контекста, например наличие четкой артикуляции при разговоре;
- различные помехи (отражения звука, искажение микрофона, фоновый шум).

Отличием распознавания слитной речи от, например, отдельных команд или подготовленной речи, являются различные сбои в произношении [9] [10]. Очень сложно говорить гладко (не сбиваясь) и красиво оформлять свои мысли (четко и ровно составлять предложения), поэтому можно сказать, что основная особенность слитной речи - это сбивчивость, наличие повторений, пауз, слов в упрощенной форме (разговорный стиль) [11].

Такие особенности зачастую являются препятствием для обработки речи техническими средствами, так как уловить особенности разговорной речи человека довольно сложно машине, поэтому необходимо либо разработать метод, на основе которого машина научиться распознавать речь человека, либо составлять сверхбольшой словарь слов или звуков, что довольно затратно по памяти [11].

## 1.3 Нарушения и дефекты речи

У людей с дефектами речи помимо выше описанных особенностей есть и другие, не менее важные, поэтому специальные системы распознавания речи должны также распознавать разные виды нарушения устной речи.



Существует несколько классификаций нарушений:

- клинико-педагогическая классификация.
- психолого–педагогическая классификация.

В данной работе будут рассмотрены нарушения клинико-педагогической классификации [12].

Виды нарушения устной речи.

## 1. Нарушения внешнего оформления устной речи.

- (a) Дисфония – отсутствие голоса или расстройство речи вследствие патологических изменений голосового аппарата: происходят различные изменения и нарушения в силе и тембре, выражающееся в охриплости, слабости голоса.
- (b) Брадилалия – медленный темп речи вследствие поражения головного мозга. Речь сильно замедляется, становится нечеткой, растягиваются гласные.
- (c) Тахилалия – быстрый темп речи, часто сопровождающийся повторением или пропуском слов, незамеченным говорящим.
- (d) Заикание – нарушение речи, которое характеризуется частым повторением или пролонгацией звуков, слогов, слов, частыми остановками или нерешительностью в речи, разрывающее её ритмическое течение.
- (e) Дислалия – нарушение звукопроизношения при нормальном слухе и нормальной иннервации речевого аппарата, которое проявляется в заменах, искажениях и смещениях звуков родной речи.
- (f) Ринолалия – нарушение произносительной стороны речи или тембра голоса, обусловленное анатомо-физиологическим поражением речевого аппарата: струя воздуха проходит не в ротовую, а в носовую полость, в которой происходит резонанс.
- (g) Дизартрия – нарушение произносительной стороны речи вследствие поражения центральной нервной системы.

## 2. Нарушения структурно-семантического оформления.

- (а) Алалия – полное отсутствие или недоразвитие речи у детей при нормальном слухе и первично сохранном интеллекте.
- (б) Афазия – речевое расстройство уже сформировавшейся речи. Причинами могут быть перенесенные черепно-мозговые травмы, инфекционные заболевания нервной системой.

Также стоит отметить, что нарушения речи могут встречаться в комплексе, например: заикание и дизартрия.

Выше перечисленные нарушения также являются препятствием для обработки речи техническими средствами. В разделе 1.5 будет приведен эксперимент из статьи [13], в котором показано, что машина иногда не понимает, что говорит человек с нарушениями речи.

## 1.4 Методы извлечения речевых характеристик

В системах распознавания речи одну из главных ролей играет извлечение признаков (частотной характеристики), при этом характеристики сигналов возбуждения чаще всего отбрасываются.

Извлечение признаков – это процесс удаления ненужной и избыточной информации и сохранение только полезной информации. Цель такого действия состоит в том, чтобы определить набор свойств (параметры), путем обработки формы сигнала поступившего на вход системе. Извлечение признаков включает процесс преобразования речевых сигналов в цифровую форму и измерение важных характеристик сигнала, например, энергии или частоты, и дополнение этих измерений значимыми производными измерениями. [14] [15].

Методы извлечения признаков удобно применять при обработке речи с неправильным произношением звуков, так как их можно адаптировать, извлекая нужные параметры, которые потребуются в дальнейшем.

### 1.4.1 Линейно-предсказывающее кодирование

Линейно-предсказывающее кодирование (англ. Linear prediction coding (LPC)).

Принцип метода линейного предсказания состоит в том, что участок речевого сигнала можно аппроксимировать линейной комбинацией предыдущих участков сигнала. Предполагается, что речь создается возбуждением линейного изменяющегося во времени фильтра (речевого тракта) случайным шумом для невокализованных речевых сегментов или последовательностью импульсов для голосовой речи [16].

Процесс речеобразования описывается линейной системой с переменными параметрами и передаточной функцией <sup>1</sup> (1.1).

$$H(z) = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}}, \quad (1.1)$$

где  $G$  – коэффициент усиления,  $a_k$  – коэффициент предсказания,  $p$  – порядок линейного предсказания.

Зависимость  $n$ -го отсчета речевого сигнала  $s(n)$  от сигнала возбуждения  $u(n)$  выражается в виде (1.2)

$$s(n) = \sum_{k=1}^p a_k s(n-k) + Gu(n), \quad (1.2)$$

Суть данного алгоритма заключается в нахождении *линейных коэффициентов предсказания* по речевому сигналу с минимизацией погрешности.

Недостатком этого метода является то, что он сильно зависит от точности произношения.

---

<sup>1</sup>Передаточной функцией называется отношение изображения выходного воздействия к изображению входного при нулевых начальных условиях.

## 1.4.2 Мэл-частотные кепстральные коэффициенты

Мэл-частотные кепстральные <sup>2</sup> коэффициенты (англ. Mel frequency Cepstral Coefficient (MFCC)).

Мел <sup>3</sup>-частотный анализ представляет частоты речи с позиции психо-акустического параметра слуха – высоты тона. Высота тона определяет, насколько высоким или низким кажется тон слушателю. Связь между частотой звука и его высотой представлена на рисунке 1.2. [16] [17]

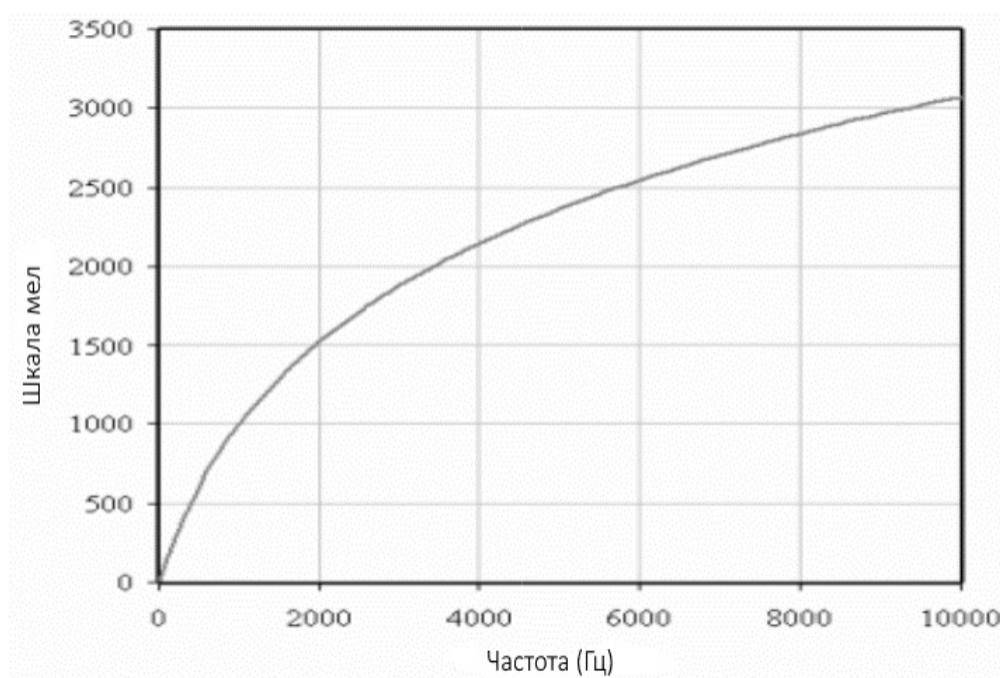


Рисунок 1.2 – Связь между частотой звука и его высотой

Положительные моменты при использовании кепстральных коэффициентов:

- спектр проецируется на специальную Мел-шкалу, позволяя выделить наиболее значимые для восприятия человеком частоты;
- количество вычисляемых коэффициентов может быть ограничено любым значением.

---

<sup>2</sup>Кепстр (cepstrum) — это результат дискретного косинусного преобразования от логарифма амплитудного спектра сигнала.

<sup>3</sup>Мел – единица высоты звука.

### 1.4.3 Кепстральные коэффициенты на основе линейного предсказания

Кепстральные коэффициенты на основе линейного предсказания (англ. Linear prediction cepstral coefficient (LPCC)).

Одна из причин использования данного метода в задачах распознавания речи состоит в том, что кепстр описывает огибающую спектра сигнала в сжатом виде. [16]

Алгоритм можно разделить на несколько этапов:

1. Речевой сигнал проходит предобработку фильтром, который усиливает высокие частоты спектра, которые в свою очередь уменьшаются в процессе воспроизведения речи.
2. Сигнал делится на одинаковые последовательные перекрывающиеся временные участки – фреймы. С помощью преобразования Фурье для каждого участка находится среднее значение частот.
3. Производятся математические вычисления, а именно берется логарифм от полученного ранее значения и выполняется дискретное косинусное преобразование.

### 1.4.4 Дискретное вейвлет-преобразование

Дискретное вейвлет-преобразование (Discrete Wavelet Transform (DWT)).

Для наиболее информативного анализа сложных реальных сигналов необходима обработка как по частотным, так и по временным характеристикам, а также достоверное представление уровней детализации для обнаружения закономерностей.

Идея применения вейвлетов состоит в многомасштабной обработке сигнала, т. е. в анализе сигнала в разном увеличении с разной степенью детализации. [17]

Недостатком вейвлет преобразований можно считать их относительную сложность расчетов.

## 1.5 Рассмотрение эксперимента

В статье [13] проводится эксперимент работы трех платформ САРР (Amazon, Google, IBM) для групп людей: с нейродегенеративными (медленно прогрессирующие, наследственные или приобретенные заболевания нервной системы, ведущими к различным симптомам – к деменции <sup>4</sup>, нарушению движения, а в следствие чего к нарушениям речи) заболеваниями и здоровых.

Записи чтения текста были расшифрованы с помощью САРР и вручную, затем сравнены. Точность расшифровки измерялась как доля верно распознанных слов.

Результат эксперимента ожидаем: точность расшифровки САРР для здоровых людей выше, чем для людей с заболеваниями (рассматривались 2 группы людей, с различными заболеваниями: с рассеянным склерозом и с атаксией Фридрейха <sup>5</sup>). При этом при увеличении продолжительности болезни, точность расшифровки САРР снижалась.

Часть результата приведена на рисунке 1.3

Обозначения:

- Группа 1 – группа здоровых людей;
- Группа 2 – группа людей с рассеянным склерозом;
- Группа 3 – группа людей с заболеванием "Атаксия Фридрайха".

Как видно из эксперимента, системы распознавания речи работают с ошибками для людей с нейродегенеративными заболеваниями.

---

<sup>4</sup>Деменция — это синдром, возникающий при поражении головного мозга и характеризующийся нарушениями в когнитивной сфере (восприятие, внимание, узнавание, память, интеллект, речь).

<sup>5</sup>Атаксия Фридрейха — генетическое заболевание, связанное с нарушением транспорта железа из митохондрий и протекающее с преимущественным поражением клеток центральной и периферической нервной системы.

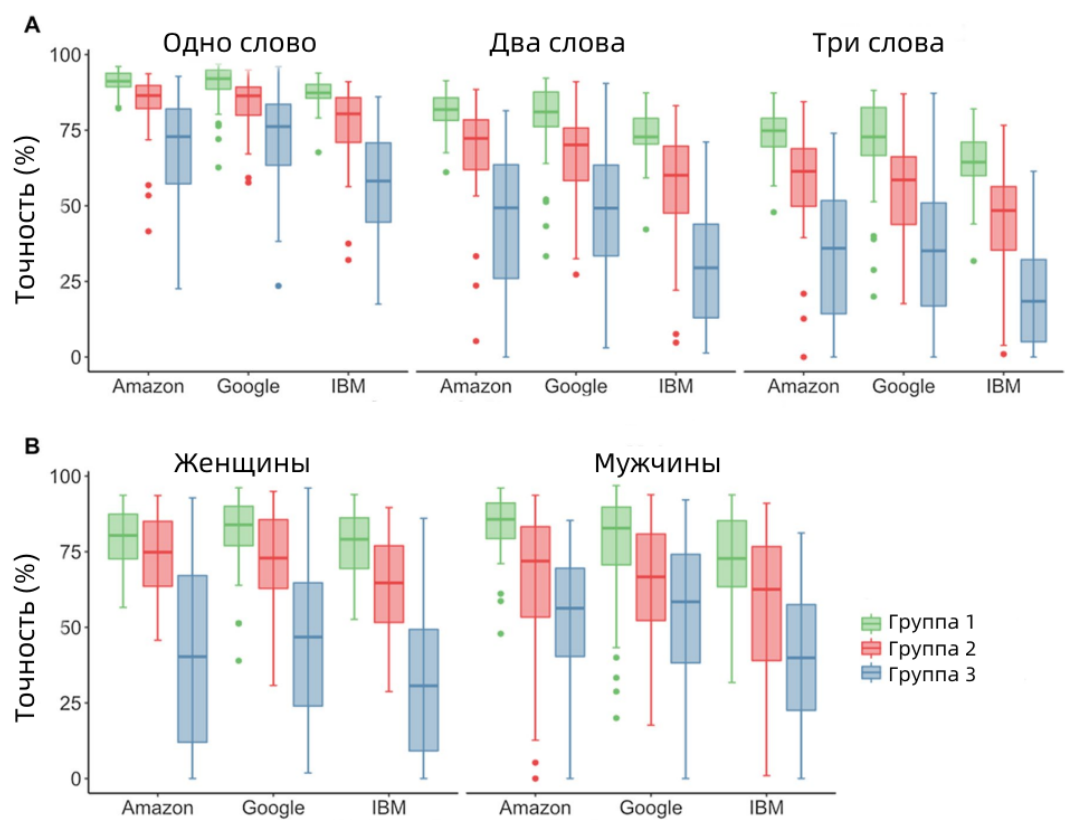


Рисунок 1.3 – Результат эксперимента

# Литература

- [1] Bhuvaneshwari Jolad D. R. K. Different feature extraction techniques for automatic speech recognition // International Journal of Engineering Sciences, Research Technology. 2018. February. P. 181–188.
- [2] Shreya Narang M. D. G. Speech Feature Extraction Techniques // International Journal of Computer Science and Mobile Computing. 2015. March. P. 107 – 114.
- [3] И.Б. Тампель А.А. Карпов. Автоматическое распознавание речи. Университет ИТМО, 2010. с. 138.
- [4] Федосин С.А. Еремин А. Ю. Классификация систем распознавания речи.
- [5] Д.Н. Бабин И.Л. Мазуренко А.Б. Холоденко. О перспективах создания системы автоматического распознавания слитной устной русской речи. С. 10–15.
- [6] А.П.Зубаков. Фурье и Вейвлет-преобразования в проблемах распознавания речи // Вестник ТГУ. 2010. т.15, вып.6. С. 1893–1899.
- [7] А.К.Алимурадов А.Ю.Тычков А.П.Зарецкий А.П.Кулешов. Способ определения кепстральных маркеров речевых сигналов при психогенных расстройствах // Труды МФТИ. 2017. Том 9, №4. С. 201–214.
- [8] Taabish G. A. S. A Systematic Analysis of Automatic Speech Recognition: An Overview // International Journal of Current Engineering and Technology. 2014. E-ISSN 2277 – 4106, P-ISSN 2347 - 5161. P. 1664–1675.
- [9] R. P. K. R. Continuous Speech Recognition // Asian Journal of Computer Science And Information Technology. 2014. 62 - 66. P. 62–66.
- [10] В.О. Верховданова А.А. Карпов. Моделирование речевых сбоев в системах автоматического распознавания речи. С. 10–15.
- [11] А.А. Леонович. Проблемы распознавания слитной речи // Цифровая Обработка Сигналов. 2007. No4. С. 1664–1675.



- [12] Речевые нарушения [Электронный ресурс]. Режим доступа: <https://lena.spb.ru/rechevye-narusheniya.html> (дата обращения: 11.12.2021).
- [13] S. B. G. V. Feature Extraction Techniques for Speech Recognition // International Journal of Speech Technology. 2021. 24:771–779. P. 771–780.
- [14] Praphulla A. Sawakare R. R. Speech Recognition Techniques // International Journal of Scientific, Engineering Research. 2015. Volume 6, Issue 8, August-. P. 1664–1675.
- [15] Kishori R. R. R. Feature Extraction Techniques for Speech Recognition // International Journal of Scientific, Engineering Research. 2015. Volume 6, Issue 5, Ma. P. 143–148.
- [16] А.В.Судьенкова. Обзор методов извлечения акустических признаков речи в задаче распознавания диктора // Сборник научных трудов НГТУ. 2019. № 3–4 (96). С. 139–164.
- [17] Первушин Е.А. Обзор основных методов распознавания дикторов // Математические структуры и моделирование. 2011. вып. 24. С. 41–54.