Санкт-Петербургский государственный университет

**СИЛКОВ Сергей Владимирович**

**Выпускная квалификационная работа**

**Автоматическое распознавание английских заимствований в русской речи (на примере сферы IT)**

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5892.2022 «Искусственный интеллект в моделировании речевой деятельности»

Научный руководитель:

к.ф.н., доцент.,

Кафедра фонетики и методики

преподавания иностранных языков,

Евдокимова Вера Вячеславовна

Рецензент:

старший научный сотрудник, Лаборатория речевых и многомодальных интерфейсов, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», Санкт-Петербургский институт информатики и автоматизации Российской академии наук,
Кипяткова Ирина Сергеевна

Санкт-Петербург

2024

**Содержание**

[Введение 3](#_heading=h.30j0zll)

[**ГЛАВА I. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ И АНАЛИЗ АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ В РУССКОЙ РЕЧИ 5**](#_heading=h.1fob9te)

[1.1. Основы автоматического распознавания речи 5](#_heading=h.3znysh7)

[1.1.1. Концепт и основные компоненты автоматического распознавания речи 5](#_heading=h.2et92p0)

[1.1.2. Эволюция автоматического распознавания речи 9](#_heading=h.tyjcwt)

[1.1.3. Проблемы распознавания многоязычной речи 16](#_heading=h.3dy6vkm)

[1. 2. Жаргон и английские заимствований в русской речи сферы IT 17](#_heading=h.1t3h5sf)

[1.2.1. Специфика жаргона IT 17](#_heading=h.4d34og8)

[1.2.2. Адаптация английских заимствований и их типы 18](#_heading=h.2s8eyo1)

[1.3. Выводы к главе I 19](#_heading=h.17dp8vu)

[**ГЛАВА II. СОСТАВЛЕНИЕ КОРПУСА АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ И РАЗРАБОТКА МЕТОДИК УЛУЧШЕНИЯ РАСПОЗНАВАНИЯ АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ 21**](#_heading=h.3rdcrjn)

[2.1. Методология проведения исследования 21](#_heading=h.26in1rg)

[2.2. Корпус английских заимствований 23](#_heading=h.lnxbz9)

[2.2.1. Методология сбора и инструменты обработки данных корпуса 23](#_heading=h.35nkun2)

[2.2.3. Структурирование и аннотация корпуса 28](#_heading=h.1ksv4uv)

[2.2.4. Анализ корпуса данных 29](#_heading=h.44sinio)

[2.3. Анализ эффективности системы Yandex SpeechKit в распознавании англицизмов на различных уровнях их освоенности 31](#_heading=h.2jxsxqh)

[2.4. Дообучение модели Whisper Open AI на корпусе английских заимствований "English Borrowings in Russian IT Speech" 37](#_heading=h.z337ya)

[2.4.1. Автоматическое распознание английских заимствований без дообучения модели Whisper Open AI 37](#_heading=h.1y810tw)

[2.4.2. Описание кода дообучения модели Whisper Open AI 38](#_heading=h.4i7ojhp)

[2.4.3. Процесс дообучения 41](#_heading=h.2xcytpi)

[2.4.4. Оценка и анализ результатов дообучения 42](#_heading=h.1ci93xb)

[2.5. Выводы к главе II 43](#_heading=h.3whwml4)

[**Заключение 44**](#_heading=h.2bn6wsx)

[Список литературы 47](#_heading=h.qsh70q)

[**Приложения 54**](#_heading=h.3as4poj)

# Введение

Актуальность автоматического распознавания речи (Automatic Speech Recognition – ASR) продолжает расти с развитием технологий и увеличением потребности в эффективной обработке больших объемов информации. Особое внимание следует уделить распознаванию русской речи с английскими заимствованиями, так как это актуально для международных и многоязычных сфер деятельности, таких как информационные технологии (Information Technologies – IT). Применение жаргона и английских терминов в русскоязычном общении в этой сфере является обыденным, что создает дополнительные трудности для систем ASR.

**Целью** данного исследования является разработка методик автоматического распознавания русской речи с английскими заимствованиями, с особым акцентом на IT-жаргон. Это особенно актуально в свете последних разработок компании Zoom, которая представила новый инструмент для автоматического резюмирования встреч и совещаний. Данный инструмент позволяет пользователям получать автоматические сводки встреч, что упрощает обработку информации и повышает производительность команды. Технология автоматического распознавания речи лежит в основе функционала автоматического создания итогов по совещаниям. Она обрабатывает аудио- и видеозаписи совещаний, преобразуя устные высказывания в текст. После этого, с применением алгоритмов искусственного интеллекта, происходит анализ текста, в ходе которого система выявляет главные моменты и важные действия, озвученные участниками. Однако, было отмечено, что качество распознавания русской речи с английскими заимствованиями может быть недостаточным, что подчеркивает важность дальнейших исследований в этом направлении для улучшения качества распознавания англицизмов и обеспечения более эффективного автоматического распознавания речи.

 **Задачи** данного исследования:

1. Анализ существующих научных статей для определения основ автоматического распознавания речи и проведение оценки существующих подходов к распознаванию речи и обработки английских заимствований в таких системах;
2. Создание корпуса английских заимствований: сбор и систематизация примеров англицизмов, используемых в русской речи в контексте IT-сферы;
3. Дообучение модели Open AI Whisper: разработка и реализация процесса дообучения модели Open AI Whisper с использованием корпуса англицизмов для повышения точности распознавания;
4. Тестирование и оценка эффективности: Проведение серии экспериментов для оценки улучшений в точности и надежности ASR после интеграции корпуса и дообучения модели;

Исследование, основанное на анализе IT-жаргона и английских заимствований в русской речи, изучает их влияние на эффективность систем автоматического распознавания речи (ASR). Ожидается, что выводы этого исследования не только улучшат работу систем распознавания русской речи, но и расширят знания о вызовах автоматического распознавания многоязычной и профессиональной речи в более широком контексте.

# ГЛАВА I. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ И АНАЛИЗ АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ В РУССКОЙ РЕЧИ

## 1.1. Основы автоматического распознавания речи

### 1.1.1. Концепт и основные компоненты автоматического распознавания речи

Автоматическое распознавание речи (ASR, от англ. Automatic Speech Recognition) – это технологический процесс, преобразующий аудиосигнал речи в текстовую форму. Системы ASR анализируют звуковые волны, создаваемые при говорении, и преобразуют их в слова и фразы, которые затем могут быть использованы для различных приложений [IBM: [сайт]. URL: https://www.ibm.com/cloud/learn/speech-recognition].

Технология автоматического распознавания речи впервые появилась в 1952 году с разработкой программы Audrey от Bell Labs, которая могла распознавать простые числа. В 1970-х годах достигнут прогресс благодаря внедрению скрытых марковских моделей, которые определяют вероятные слова для расшифровки, используя вероятностные функции. С 2010 по 2020 год технология ASR значительно усовершенствовалась, став широко использоваться в повседневной жизни для голосовых поисков, отправки текстовых сообщений и взаимодействия с голосовыми помощниками, такими как Amazon Alexa и Apple Siri. Эти улучшения способствовали разработке новых приложений, таких как системы автоматического управления для умных домов. [Deepgram: [сайт]. URL: <https://deepgram.com/learn/the-history-of-automatic-speech-recognition>; arXiv: [сайт]. URL: <https://arxiv.org/abs/2108.00084>].

Технология ASR нашла широкое применение в различных областях. Некоторые из них включают:

1. Голосовые интерфейсы. ASR позволяет создавать голосовые интерфейсы для виртуальных помощников, улучшая взаимодействие пользователя с технологическими устройствами и сервисами;
2. Расшифровка содержания. Автоматическая расшифровка аудио и видеозаписей с помощью ASR экономит время и усилия, необходимые для ручной расшифровки;
3. Командный ввод. ASR обеспечивает возможность управлять устройствами и приложениями голосом, что может быть полезно в ситуациях, когда ручной ввод невозможен или неудобен;
4. Автоматический перевод. В комбинации с технологией машинного перевода ASR может обеспечивать мгновенный перевод голосовой речи;
5. Идентификация говорящего. ASR может быть использовано для идентификации говорящего на основе характеристик его голоса;
6. Доступность. Технология помогает людям с ограниченными возможностями, предоставляя альтернативные способы взаимодействия с технологией;
7. Колл-центры и обслуживание клиентов. ASR может быть использовано для расшифровки и анализа разговоров в колл-центрах для улучшения качества обслуживания.

ASR проявляет свой потенциал как эффективный инструмент для оптимизации пользовательских интерфейсов и предоставления новых возможностей в различных отраслях. Непрерывные исследования и разработки в сфере ASR расширяют функционал этой технологии, включая применение глубоких нейронных сетей, которые улучшают точность и эффективность процесса распознавания речи. [NVIDIA: [сайт]. URL: <https://developer.nvidia.com/blog/an-easy-introduction-to-speech-ai/>; IBM: [сайт]. URL: <https://www.ibm.com/cloud/learn/speech-recognition>].

 ASR состоит из двух главных компонентов – акустический и языковой.

Акустическая модель

Акустическое моделирование занимается анализом необработанных речевых аудиосигналов, обычно на уровне отдельных символов или слогов, предсказывая, каким фонемам или другим языковым элементам соответствуют эти аудиосигналы. Эти модели устанавливают связи между звуковым сигналом и фонемами или другими лингвистическими единицами, формирующими речь. Обычно такие сложные модели строятся на основе скрытых марковских моделей (HMM) или искусственных нейронных сетей (NN).

Модель обучается на большом объеме обучающих данных, которые используются для оценки параметров модели. Для эффективной работы систем автоматического распознавания речи (ASR) критически важны акустические модели, которые включают в себя два основных компонента. Первый компонент — это глубокие нейронные сети, задача которых заключается в анализе уровней звуковых волн. Второй компонент — скрытые марковские модели, которые используются для преобразования результатов анализа в последовательные прогнозы. Ошибки в прогнозировании последовательностей фонем могут привести к каскаду ошибок в других частях системы, что негативно сказывается на правильности интерпретации речевых данных [Bhatt и др., 2020].

Одним из традиционных и широко используемых методов акустического моделирования является HMM. HMM использовались для моделирования временной изменчивости речи и представления вероятностной природы речевого сигнала. Например, исследование (исследование инвариантности дивергенции и ее применение к распознаванию речи от Y. Qiao и N. Minematsu) показало, что определенные инвариантные структуры превосходят HMM в акустическом моделировании речевых звуков. Другое исследование подчеркнуло осуществимость использования HMM для автоматического обнаружения подводных биофонических звуков, подчеркивая универсальность HMM в различных акустических сценариях [Qiao и др., 2010; Putland и др., 2018].

С появлением глубокого обучения глубокие нейронные сети (DNN) стали выделяться в акустическом моделировании. DNN показали значительные улучшения по сравнению с традиционными HMM, особенно при их комбинировании в гибридной структуре DNN-HMM. Исследования показывают, что улучшенные методы акустического моделирования с использованием DNN приводят к улучшению непрерывного распознавания речи как в чистых, так и в шумных условиях [Dighe и др., 2016].

Языковая модель

Моделирование языка является ключевым аспектом обработки естественного языка и вычислительной лингвистики. Оно включает в себя прогнозирование следующего слова в последовательности на основе предшествующих слов. Цель заключается в том, чтобы захватить структуру и закономерности языка, которые могут быть использованы в различных приложениях, таких как распознавание речи, машинный перевод и генерация текста.

Одной из традиционных техник в моделировании языка является использование моделей N-грамм. N-грамма - это непрерывная последовательность N элементов из данной выборки текста или речи. Модели N-грамм прогнозируют следующее слово в последовательности на основе предыдущих N-1 слов. Эти модели были основополагающими в моделировании языка и показали свою эффективность, особенно когда дополняются другими моделями. Кроме того, существуют синтаксические модели N-грамм, которые не только улучшают стандартные модели N-грамм, но и превосходят связанные синтаксические языковые модели [Bellegarda, 2000; Sennrich, 2015].

С появлением техник глубокого обучения рекуррентные нейронные сети (RNN) получили широкое распространение в области языкового моделирования. Эти сети, благодаря своей способности захватывать долгосрочные зависимости в текстовых данных, идеально подходят для анализа последовательностей на естественном языке. Применение вариаций RNN, таких как долгосрочная краткосрочная память (LSTM) и управляемые рекуррентные блоки (GRU), позволило значительно улучшить воспроизведение сложных языковых структур в долгосрочной перспективе [Takahashi и др., 2019].

Лексикон – еще один важный компонент в моделировании языка. Это коллекция слов и их значений, и он играет ключевую роль в понимании и создании языка. Интеграция лексикона с языковыми моделями может улучшить производительность модели, особенно в задачах, требующих глубокого понимания значений и отношений слов [Rasipuram и др., 2015].

### 1.1.2. Эволюция автоматического распознавания речи

С момента начала разработки автоматического распознавания речи был достигнут значительный прогресс. Ранние системы сыграли ключевую роль в создании основ для дальнейшего развития этой технологии. В 1960-х годах была разработана одна из таких систем — IBM Shoebox. Эта система обладала способностью распознавать числа и математические команды, включая операции, такие как "плюс" и "итого". Shoebox использовала эти данные для решения математических задач, что представляло собой значительный шаг в направлении развития интерактивных систем распознавания речи [Валуйцева, 2019​].

После IBM Shoebox, в 1971 году, IBM создала систему автоматической идентификации вызовов, которая позволяла инженерам взаимодействовать с устройствами с помощью речи и получать устные ответы. Это новшество стало началом для технологий распознавания речи. Значительный рубеж был достигнут в 1976 году с разработкой системы Harpy в Карнеги-Меллоне после пятилетних исследований, финансированных DARPA. Harpy сделала большой шаг вперед, поскольку могла понимать 1,011 слов, что значительно превышало возможности предшественников​ [Сondecosoftware: [сайт]. URL: [<https://www.condecosoftware.com/blog/a-history-of-voice-recognition-technology>]​.

Разработка Harpy также впервые применила скрытые марковские модели (HMM), что значительно способствовало развитию технологий ASR в 1980-х годах. Это развитие позволило IBM предложить инструменты для преобразования речи в текст и провести бета-тестирование технологий расшифровки, что, в свою очередь, стало отправной точкой для коммерческих приложений распознавания речи, таких как Tangora и Dragon Naturally Speaking​3​. Значительно, что система Harpy преодолела порог точности в 95%, став первой системой распознавания речи, читаемой машинами, и установившей предпосылки для создания первой AI-управляемой системы распознавания речи в 1976 году​ [Condecosoftware : [сайт]. URL: <https://www.condecosoftware.com/blog/a-history-of-voice-recognition-technology>].

С 1976 года в области ASR произошли значительные изменения, особенно благодаря внедрению методов машинного обучения и глубокого обучения. Эти достижения не только повысили точность систем распознавания речи, но и расширили их применение. Недавние исследования по ASR:

Глубокое обучение сделало значительные шаги в распознавании речи, но исследователи, такие как Маркус (2018), предполагают, что оно должно дополняться другими техниками для достижения искусственного общего интеллекта. Это мнение находит отклик в научном сообществе, которое продолжает исследовать синергию между глубоким обучением и традиционными методами [Marcus, 2018].

В области визуального распознавания речи работа Ма (2022) показывает, что, хотя большие наборы данных способствовали улучшениям, разработка лучших моделей и стратегий аугментации данных также важна. Аналогично, Хаеб-Умбах (2020) продемонстрировал, что сочетание глубокого обучения с традиционной обработкой сигналов может привести к эффективным решениям для распознавания речи на расстоянии [Haeb-Umbach и др., 2020].

Докуз и соавт. (2021) улучшили системы ASR на основе глубокого обучения, введя стратегии выбора мини-пакетов образцов, учитывающие пол и акцент, что привело к лучшей производительности модели. Кроусон и соавт. (2020) предполагают, что глубокое обучение может преобразить интерпретацию аудиограмм, сделав ее автоматической и точной [Dokuz и др., 2021; Crowson и др., 2020].

Jermsittiparsert и соавт. (2020) разработали модель для распознавания эмоций, которая достигла точности до 94,21%, демонстрируя потенциал глубокого обучения в понимании эмоционального контекста в речи [Jermsittiparsert и др., 2020].

Впереди, Crowson и соавт. (2016) предсказывают, что глубокое обучение будет основной движущей силой в продвижении систем распознавания устной речи. Однако авторы подчеркивают компромисс между необходимостью инженерии признаков в традиционном машинном обучении и более длительным временем обучения, требуемым для моделей глубокого обучения [Crowson, 2016].

Huang и его коллеги в 2016 году продемонстрировали, что глубокие нейронные сети значительно повышают точность распознавания речи, особенно когда используется словарь большого размера, по сравнению с традиционными скрытыми марковскими моделями (HMM).[Huang и др., 2016].

Вычислительные требования глубокого обучения вызывают определённые проблемы, как отмечают Thompson и соавт. (2020), указывая на необходимость более эффективных методов, которые могут появиться благодаря достижениям в глубоком обучении или альтернативных подходах машинного обучения. Joy и соавт. (2017) улучшили акустические модели для дизартрической речи в базе данных TORGO и подчеркнули роль искусственных нейронных сетей (ИНС) в таких приложениях речи, как LVCSR и автоматическое распознавание языка [Thompson и др., 2020; Joy и др., 2017]

Эти достижения отражают тенденцию к созданию более сложных, тонких и способных систем ASR, движимых мощью глубокого обучения.

Ограничение работы систем ASR

Несмотря на значительные достижения в технологии автоматического распознавания речи, она все еще сталкивается с важными проблемами, которые ограничивают её применение в более широком спектре областей. Одной из главных трудностей является учёт человеческого фактора, который отстаёт от технологического прогресса. Noyes и соавторы подчеркнули, что несмотря на потенциальные преимущества ASR для людей с ограниченными возможностями, прогресс затруднен из-за недостатка исследований в области пользовательского опыта и практического применения. Это усугубляется тем, что технология показывает недостаточную эффективность для людей с нейродегенеративными заболеваниями, как отметили Шульц и другие в 2021 году. Это подчёркивает необходимость разработки систем автоматического распознавания речи, способных адаптироваться к различным речевым паттернам, включая те, которые изменены состоянием здоровья.[Noyes, 1989; Schultz, 2021].

Сложности речи, такие как темп и плотность информации, также представляют существенные вызовы для систем ASR. Исследование Родеро в 2016 году выявило оптимальный темп речи для распознавания, что предполагает, что системы ASR должны быть достаточно гибкими, чтобы адаптироваться к естественным колебаниям человеческой речи. Кроме того, вопрос обнаружения и исправления ошибок критичен, исследователь подчеркивая слабыестороны современной технологии, в то же время выделяет потенциал для значительных улучшений, если будут реализованы надежные механизмы коррекции ошибок [Rodero, 2016].

Языковой барьер остается еще одним препятствием для технологии ASR. Гласс в 2012 году подчеркнул важность создания систем автоматического распознавания речи, способных обучаться и адаптироваться для понимания различных языков и диалектов, что является необходимым условием для их глобального применения. В 1985 году Уилпон указал на множество практических проблем, возникающих при внедрении систем автоматического распознавания речи, таких как шум окружающей среды и ограничения оборудования, которые часто не выявляются до начала их использования. [Wilpon и др., 1986].

Эти исследования подчеркивают многогранную природу вызовов, с которыми сталкивается технология ASR. Несмотря на достигнутый прогресс, очевидна необходимость продолжения исследований и разработок с целью сделать ASR более доступным, точным и удобным для широкого круга пользователей и приложений.

Open AI Whisper: Прогресс в области самостоятельного распознавания речи и методики дообучения

Исследование, проведенное Peng в 2023 году, подчеркивает важные достижения модели GPT-4 в области текстовой генерации, используя методы глубокого обучения. Описываемая как "Generative Pre-trained Transformer 4", GPT-4 функционирует как сложная система, обученная на обширном наборе текстовых данных. Ее эффективность в задачах анализа и генерации текста выделяет ее как ценный инструмент для адресации различных вызовов в текстовой обработке [Peng, 2023].

Первоначальное обучение GPT-4 происходит через анализ обширных текстовых данных в неуправляемом режиме, что способствует адаптации модели к языковым контекстам. В процессе дополнительного развития модели применяются новаторские методы самообучения, включая контрастные задачи. Эти методы помогают модели эффективно различать важные и второстепенные текстовые элементы, значительно улучшая её способность к обучению [Liu и др., 2024].

Дообучение модели GPT-4, проведенное на выбранном корпусе данных, целенаправленно улучшает функционирование модели в заданных условиях, минуя этап начального обучения на широких общедоступных данных. Процесс включает несколько ключевых этапов:

1. *Выбор специализированного набора данных*: необходимо точно определить и подготовить данные, которые соответствуют уникальным требованиям задачи или языковым особенностям, для последующего дообучения модели;

2. *Регулировка параметров обучения:* настройка таких параметров, как скорость обучения, количество эпох и размер партии данных, для оптимизации процесса дообучения, адаптированного под конкретный набор данных;

3. *Мониторинг и оценка результатов:* использование метрик, таких как точность, полнота и F1-мера, для оценки эффективности модели после дообучения, что позволяет убедиться в корректности адаптации модели к заданным условиям;

4. *Итерационная оптимизация:* непрерывная корректировка и детальная настройка модели на основе анализа результатов для дальнейшего повышения ее эффективности [Baktash, 2023].

Для улучшения системы автоматического распознавания русской речи, содержащей элементы английского языка, выбор модели OpenAI Whisper обусловлен рядом ключевых преимуществ:

1. *Многоязычная поддержка*. Разработанная как многоязычная платформа ASR, Whisper идеально подходит для обработки текстов, содержащих элементы разных языков. Это делает ее особенно полезной в средах, где присутствуют различные языковые структуры, что подтверждено исследованиями, демонстрирующими улучшенную производительность многоязычных моделей.

2. *Эффективность дообучения на ограниченном объеме данных.* Whisper способна эффективно адаптироваться даже при использовании сравнительно малых объемов данных. Это критически важно, когда доступные данные ограничены или необходима быстрая адаптация к специфическим требованиям (Hu и др., 2023).

Использование модели Whisper значительно ускоряет процесс разработки систем ASR и повышает их точность при работе с техническими терминами и специализированным сленгом. Эти качества делают Whisper отличным выбором для проектов по автоматическому распознаванию русской речи, содержащей английские заимствования. Эффективность этой модели в многоязычных условиях и возможность адаптации с использованием ограниченных данных подчеркивают ее значимость для улучшения распознавания речи с английскими заимствованиями.

### 1.1.3. Проблемы распознавания многоязычной речи

В эпоху цифровых технологий ASR играет ключевую роль в облегчении взаимодействия человека с машиной. Однако, несмотря на значительные достижения, системы ASR сталкиваются с серьезными проблемами, особенно при работе с многоязычными данными, такими как русский и английский языки.

Переключение между разными языками в одной системе ASR представляет собой значительную проблему. Вайбель и соавт. (1997) подчеркивают сложности, связанные с переходом между системами распознавания для разных языков. Это особенно важно в контексте русского и английского языков, которые не являются близкими. [Waibel и др., 1997].

Идентификация языка является ключевым аспектом в многоязычных системах ASR. Исследование Джейна и соавт. (2018) показывает, как разнообразие акцентов в английском языке может усложнить этот процесс. Они предлагают включение в модель акцентов и многомерного обучения для повышения производительности системы [Jain и др., 2018].

Разработка новых технологических решений необходима для преодоления этих проблем. Арора и соавт. (2019) исследуют использование пословного лексикона и языковой модели для улучшения распознавания слов, которые отсутствуют в словаре, в русском и английском языках. Это подчеркивает важность адаптивных систем, способных справляться с динамичной природой языка [Arora, 2019].

Многоязычное распознавание речи в системах ASR, особенно для русского и английского языков, остается сложной задачей. Проблемы включают в себя идентификацию языка, переключение между языковыми системами и обработку разнообразия акцентов. Однако, как показывают исследования Вайбеля и др. (1997), Джейна и др. (2018), и Ароры и др. (2019), продолжающиеся работы в этой области обещают более эффективные и надежные системы ASR для мировой аудитории.

## 1. 2. Жаргон и английские заимствования в русской речи сферы IT

### 1.2.1. Специфика жаргона IT

В современном мире, где технологии стремительно развиваются, язык информационных технологий (IT) претерпевает значительные изменения. Особенно это касается жаргонных элементов, которые становятся неотъемлемой частью как профессиональной, так и повседневной речи. В статье Николовой Д. "Коммуникация в сфере информационных технологий и жаргон как одно из ее средств" подробно рассматривается роль жаргона в коммуникации IT-сферы [Николов, 2015].

Отсутствие стандартизированной терминологии в IT-сфере приводит к распространению компьютерного жаргона, который заменяет традиционные средства вербальной и невербальной коммуникации. Этот аспект жаргона выявлен в статье как критически важный, поскольку жаргон не только обогащает язык, но и отражает культурные и социальные изменения. Особое внимание уделяется детабуизации инвективной лексики и использованию варваризмов, которые являются важными тенденциями в развитии русского и болгарского языков.

Важным аспектом является и различие в использовании терминов "жаргон" и "сленг". В контексте компьютерного или интернет-жаргона некоторые исследователи рассматривают интернет-сленг и интернет-жаргон как отдельные разновидности языка, отличные от компьютерного жаргона. Однако, термин "компьютерный жаргон" может обозначать язык интернета в целом, включая его различные локальные особенности [Николов, 2015].

Таким образом, анализ жаргона в IT-сфере позволяет не только понять его специфику, но и проследить динамику изменений в языке, вызванных технологическим прогрессом и культурными взаимодействиями.

### 1.2.2. Адаптация английских заимствований и их типы

Адаптация англицизмов в русском языке, особенно в сфере IT, представляет собой сложный процесс, включающий несколько видов адаптаций заимствований. Лингвистический энциклопедический словарь предлагает следующее определение слова «заимствование»: «Заимствование — элемент чужого языка (слово, морфема, синтаксическая конструкция и т. п.), перенесенный из одного языка в другой в результате языковых контактов, а также сам процесс перехода элементов одного языка в другой» [ЛЭС, 2002].
Рассмотрим основные типы адаптаций и примеры:

1. *Фонетическая адаптация:* процесс приспособления исходного английского произношения к фонетическим нормам русского языка. Например, слово "computer" адаптируется как "компьютер";
2. *Семантическая адаптация:* изменение или расширение значения слова в контексте русского языка. Например, "server" в английском языке означает "сервер", но в русском оно может использоваться более широко для обозначения различных видов обслуживающего оборудования;
3. *Синтаксическая интеграция:* включение англицизмов в структуру русского предложения, что может влиять на порядок слов и грамматические конструкции. Например, использование "online" как "онлайн" в различных контекстах предложения;
4. *Морфологическая адаптация:* приспособление англицизмов к морфологическим правилам русского языка, включая склонение и спряжение. Например, слово "download" превращается в "даунлоадить" [Zolotareva, 2016].

Эти процессы иллюстрируют сложность включения англицизмов в русский язык и выдвигают особые требования к системам автоматического распознавания речи, особенно в сфере IT. Они требуют продвинутых алгоритмов распознавания и обработки речи для адекватного восприятия и интерпретации этих заимствованных слов.

Английские заимствования в русском языке представляют собой интересное явление, включающее в себя различные типы заимствований из английского языка. Они классифицируются на три основные группы:

*1. Освоенные англицизмы:* эти слова полностью интегрированы в русский язык и активно используются в повседневной речи. Примеры включают слова, такие как "теннис", "менеджер", которые были полностью ассимилированы;

2. Полуосвоенные англицизмы: это заимствования, которые недавно вошли в русский язык и могут быть не полностью известны всем носителям языка. Примеры могут включать термины, такие как "логин", "хостинг";

3. Неосвоенные англицизмы: они ограниченно используются и часто связаны со специфическими областями, как окказионализмы или трансплантанты, например, профессиональный жаргон в IT [Дьяков, 2016].

Эти категории помогают понять, как англицизмы адаптируются в русском языке, и какие вызовы они представляют для систем автоматического распознавания речи.

## 1.3. Выводы к главе I

В первой главе нашего исследования был проведён тщательный анализ теоретических основ автоматического распознавания речи, а также рассмотрен вопрос заимствования английской лексики в русский язык. Ключевые аспекты первой главы:

**1. Эволюция ASR.** Было проанализировано историческое развитие технологий автоматического распознавания речи, начиная от первых экспериментов до применения передовых методов глубокого обучения. Этот обзор позволил подробно рассмотреть основные принципы, на которых строятся эти технологии;

**2. Вызовы многоязычного распознавания.** В работе были рассмотрены трудности, связанные с многоязычным распознаванием речи, включая проблемы переключения между языками и различия в языковых структурах, что критически важно для анализа русского и английского языков;

**3. Процесс адаптации англицизмов.** Был проведен анализ адаптации англицизмов в русском языке, их типологии и механизмов интеграции в речь. Было исследовано, как заимствования влияют на процессы распознавания, и разработаны подходы, которые могут улучшить их распознавание.

**4. Технологическая реализация ASR систем.** Были изучены технические аспекты систем автоматического распознавания речи, включая алгоритмы, используемые для их обучения и дообучения, а также особенности работы с многоязычными данными.

Эти аспекты сложили теоретическую базу для практической работы, направленной на улучшение процесса распознавания англицизмов в русском IT-жаргоне. В ходе исследования было подчеркнуто, что необходим комплексный подход к обучению и адаптации систем автоматического распознавания речи, особенно учитывая лингвистическое разнообразие и языковые контакты в современном мире.

# ГЛАВА II. СОСТАВЛЕНИЕ КОРПУСА АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ И РАЗРАБОТКА МЕТОДИК УЛУЧШЕНИЯ РАСПОЗНАВАНИЯ АНГЛИЙСКИХ ЗАИМСТВОВАНИЙ

## 2.1. Методология проведения исследования

Данное исследование посвящено улучшению автоматического распознавания англицизмов в русской IT речи. Был разработан обширный план работы, который включает в себя ряд этапов: от ручного сбора данных до их анализа и интерпретации. Целью данного проекта является не только создания корпуса англицизмов в русской речи, но и разработка эффективных методов для оптимизации систем автоматического распознавания речи.

Общий процесс проведения исследования:

**Ручной сбор данных:**

* Выборка контента: методика включает ручной отбор видеоматериалов из YouTube, ориентированных на IT-тематику с использованием англицизмов, таких как лекции, семинары, интервью с IT-специалистами;

**Обработка и расшифровка аудио**

* Извлечение аудио: использование программных средств для выделения аудиодорожек из видеофайлов;
* Ручная расшифровка: расшифровка аудиодорожек в текстовый формат.

**Подготовка данных для машинного обучения:**

* Форматирование данных: преобразование аудио и соответствующих расшифровок в формат, совместимый с моделью распознавания речи;
* Очистка и нормализация данных: удаление шумов, пауз и прочих несущественных элементов аудио для улучшения качества данных.

**Дообучение и тестирование модели:**

* Дообучение модели Open AI Whisper: использование Python и библиотек (datasets, transformers) для адаптации модели к специфике исследуемых данных;
* Валидация модели: оценка способности модели корректно распознавать и интерпретировать англицизмы в русской речи.

**Анализ и интерпретация результатов:**

* Статистический анализ: анализ частоты и распределения англицизмов в собранных материалах;
* Контекстуальный анализ: изучение использования англицизмов в различных контекстах IT-сферы.

**Публикация результатов:**

* Отчетность и публикации: представление результатов в научных журналах или на профильных конференциях для обсуждения и дальнейшего анализа.

В рамках проведённой работы был создан и систематически организован корпус аудиоданных под названием "English Borrowings in Russian IT Speech". Целью проекта является идентификация, анализ и каталогизация английских заимствований, используемых в русскоязычном контексте сферы IT. Проект имеет значительное прикладное и теоретическое значение для лингвистики, компьютерной лингвистики и исследований в области межкультурной коммуникации.

## 2.2. Корпус английских заимствований

### 2.2.1. Методология сбора и инструменты обработки данных корпуса

В ходе текущего исследования был создан корпус аудиоданных, названный "English Borrowings in Russian IT Speech" (Приложение 1), который включает английские заимствования в русской речи в области IT. Для сбора данных была привлечена группа исследователей, которые анализировали и отбирали видеоматериалы с платформы YouTube.

Процесс сбора аудиоданных для корпуса “English Borrowings In Russian IT Speech” представлял собой следующий процесс: каждый найденный англицизм из видеоматериалов был вырезан вместе с фразой в виде аудио фрагмента формата WAV, что обеспечивало высокое качество сохранения контекста речи. Параллельно с этим, исследователи создавали текстовые файлы с субтитрами для каждого аудиофрагмента, расшифровывая произнесенные слова и фразы. Такой подход позволил не только сохранить аутентичное звучание заимствованных терминов, но и обеспечил точное текстовое представление речи, что является критически важным для последующего анализа и обработки данных.

Сформированный таким образом корпус данных состоял из пар «аудиофрагмент — текстовый файл», где каждый элемент пары был строго сопоставлен друг с другом. Например, апгрейд\_27.wav и апгрейд\_27.txt. Это сопоставление значительно упрощает процесс анализа заимствований, позволяя автоматизировать обработку данных и обеспечивая высокую точность при идентификации и классификации английских слов и фраз в русском языковом контексте.

Такая методика сбора и организации исследовательского материала обеспечила надежную основу для последующей разработки и усовершенствования алгоритмов автоматического распознавания речи, способных эффективно определять и обрабатывать англицизмы в русской речи, что особенно актуально для профессионального общения в сфере IT.

Для формирования корпуса было просмотрено 3000 видео с YouTube, отобранных на основании тематики, которая посвящена информационным технологиям. В результате исследовательской работы было извлечено 2150 аудиозаписей с продолжительностью каждой от 5 до 8 секунд. Каждая запись сопровождается текстовым файлом с точной расшифровкой произнесенных фраз, что облегчает дальнейший анализ и обработку данных.

**Критерии отбора видео:**

 Для обеспечения качественного и целенаправленного сбора данных в рамках проекта по автоматическому распознаванию английских заимствований в русской речи, особое внимание было уделено отбору видеоматериалов с YouTube. Основным инструментом для этого послужил список ключевых слов, связанных с сферой информационных технологий и машинного обучения. Этот список помог обеспечить высокую релевантность и предметную наполненность аудиокорпуса.

Таблица 1. Список ключевых слов, использованных для поиска видеоматериалов

|  |  |
| --- | --- |
| 1. Искусственный интеллект (AI)
 | 1. Робототехника
 |
| 1. Машинное обучение (ML)
 | 1. Автоматизация
 |
| 1. Глубокое обучение (Deep Learning)
 | 1. DevOps
 |
| 1. Блокчейн
 | 1. Контейнеризация
 |
| 1. Облачные технологии
 | 1. Microservices (Микросервисы)
 |
| 1. Кибербезопасность
 | 1. Финтех
 |
| 1. Big Data
 | 1. Смарт-контракты
 |
| 1. Интернет вещей (IoT)
 | 1. Python
 |
| 1. Самоуправляемые автомобили
 | 1. JavaScript
 |
| 1. Виртуальная реальность (VR)
 | 1. Java
 |
| 1. Дополненная реальность (AR)
 | 1. C#
 |
| 1. Квантовые вычисления
 | 1. PHP
 |
| 1. 5G технологии
 | 1. Ruby…
 |

Более подробная информация о списке ключевых слов представлена в Приложении 2 под заголовком "Список ключевых слов, использованных для поиска видеоматериалов". Но также стоит отметить, что это не полный список ключевых слов и исследователи часто добавляли слова из русского языка к ключевым словам, чтобы находить более релевантные запросы.

**Инструменты использовавшиеся во время сбора корпуса аудиоданных:**

*1. YouTube - поиск видео*

YouTube является крупнейшим в мире видеохостингом, где пользователи могут загружать, просматривать, оценивать, делиться и комментировать видео. Для целей исследования, YouTube используется для поиска и просмотра видеоматериалов, связанных с IT и содержащих примеры употребления англицизмов в русской речи.

Это могут быть обучающие видео, интервью с IT-специалистами, вебинары и другие материалы на русском языке (см. Приложение 2. Список ключевых слов, использованных для поиска видеоматериалов).

*2. Python*

Разработанная программа автоматизирует процесс загрузки аудиодорожек из видео на YouTube, используя данные из Excel-файла. Этот процесс начинается с загрузки необходимых библиотек, включая `pandas` для обработки данных, `pytube` для взаимодействия с YouTube и `os` для работы с файловой системой.

Основная функция, `download\_and\_save\_audio`, принимает четыре параметра: ключевое слово, ссылку на видео, временные метки и путь сохранения. С помощью `pytube` функция загружает видео, извлекает аудиодорожку, обрабатывает каждый указанный интервал времени и сохраняет результат в формате `.mp3`. Если в процессе возникают ошибки, функция сообщает о них и возвращает статус неудачи операции.

Перед началом обработки данных удостоверяются в наличии или создают необходимую директорию для сохранения файлов. Данные из Excel файла, содержащие ключевые слова, ссылки на видео и временные метки, загружаются в `DataFrame`. Каждая строка данных обрабатывается индивидуально, и для каждой комбинации ключевого слова, ссылки и временных меток вызывается функция загрузки аудио.

Этот подход позволяет нам создать надежный и автоматизированный способ сбора аудиоданных, которые могут быть использованы в различных приложениях и исследованиях, например, в анализе мультимедийного контента или в задачах машинного обучения, где необходима обработка аудио. Это значительно упрощает процесс сбора данных и повышает эффективность работы исследователей и разработчиков, предоставляя легкий доступ к необходимым аудиоресурсам. Полный код представлен в Приложении 3. Экстрактор аудио по временным рамкам из YouTube видео корпуса “English Borrowings in Russian IT Speech”.

*3. Текстовые файлы (TXT) с расшифровкой (транскрибацией)*

Создание текстовых файлов из аудиозаписей речи, которые расшифровываются, является ключевым аспектом обработки речевых данных. Эти текстовые файлы, содержащие точную расшифровку слов, произносимых в аудиозаписях, служат метками для обучения моделей распознавания речи. В результате, модель “Whisper” может в будущем научиться сопоставлять звуковые волны с соответствующими словами и фразами, включая англицизмы, поскольку каждый аудиофайл сопровождается соответствующим текстовым файлом.

**Критерии выбора ключевых слов:** Выбор слов в этом списке англицизмов, используемых в русском языке в контексте IT, отражает широкое проникновение английского языка в техническую и бизнес-среду. Можно отметить следующие характеристики выбранных слов:

*1. Специфика IT-сферы.* Многие из этих слов, такие как "баг", "дебаггинг", "апп", "куки", "девелопер", напрямую связаны с программированием, веб-разработкой и IT-инфраструктурой. Эти термины возникли в англоязычной IT-среде и были заимствованы русским языком из-за отсутствия адекватных аналогов;

*2. Краткость и понятность.* Некоторые англицизмы используются из-за их краткости и лаконичности, например, "апдейт" вместо "обновление", "юзер" вместо "пользователь". Это делает коммуникацию более эффективной в быстро меняющейся IT-среде;

*3. Методологии и процессы. С*лова, такие как "скрам", "агайл", "девопс", отражают специфические методологии и подходы в сфере разработки программного обеспечения, которые были разработаны в англоязычном мире и приняты во всем мире без изменения названия;

*4. Влияние популярной культуры и образования.* Некоторые англицизмы, такие как "геймдев", "нейронка", стали популярными благодаря влиянию популярной культуры, медиа и специализированного образования, которое часто ведется на английском языке.

В общем, эти англицизмы стали неотъемлемой частью русского языка в контексте IT, отражая глобализацию и технологические тенденции. Они облегчают коммуникацию между специалистами, часто имеющими разное культурное и языковое происхождение, и способствуют более быстрому распространению новых технологий и концепций.

### 2.2.3. Структурирование и аннотация корпуса

Корпус "English Borrowings in Russian IT Speech" структурирован в соответствии с ключевыми тематиками, выявленными в процессе анализа видеоматериалов. Всего было создано 34 папки, каждая из которых соответствует определенному ключевому слову, включая такие аспекты, как программирование, кибербезопасность, облачные технологии и многие другие. Эта структура обеспечивает удобный доступ к материалам и их эффективное использование в исследовательских и образовательных целях.

Общая продолжительность аудиоматериалов составляет 270 минут, что предоставляет обширную базу для анализа использования английских заимствований в русской речи в контексте IT. Корпус не только способствует глубокому пониманию специфики заимствований и их функционирования в профессиональной сфере, но также может быть использован для разработки обучающих программ и инструментов для автоматического распознавания и перевода терминологии.

### 2.2.4. Анализ корпуса данных

В рамках аналитической работы по аудиокорпусу "English Borrowings in Russian IT Speech" осуществлена ранжированная категоризация уникальных терминов по частоте их употребления, начиная с наиболее часто встречающихся и завершая наименее распространенными словами (Приложение 4. Уникальные слова корпуса “English Borrowings in Russian IT Speech”).

Таблица 2. Распределение частоты упоминаний специализированных IT-терминов в аудиокорпусе "English Borrowings in Russian IT Speech"



Анализ частоты упоминания различных терминов в нашем текстовом корпусе позволяет сделать следующие вывод:

Во-первых, наличие терминов, таких как "тимлид", "айтишник" и "майнить", подчеркивает фокус на управлении командами и процессах разработки в IT. Это свидетельствует о том, что большая часть текстов в корпусе касается руководства командами разработчиков и задач, которые перед ними ставятся.

Во-вторых, термины "сабмит", "репортить", "бигдата", "спарсить" и "юзерфрендли" указывают на то, что обсуждения охватывают также технические аспекты работы в IT, включая управление данными, разработку пользовательских интерфейсов и программирование. Это подтверждает техническую направленность обсуждаемого материала.

Такой анализ позволяет сделать вывод о том, что изучаемый текстовый корпус является важным ресурсом специфики профессиональной лексики в IT-секторе и практик, характерных для этой области, что подкрепляет цели работы.

Для каждого ключевого слова был извлечен контекст из видеоматериалов. Этот процесс включал извлечение текста из фрагментов аудиоданных, в которых упоминается ключевое слово. Полученный контекст был сохранен для последующего анализа и использования в исследованиях (Приложение 5. Контекст ключевых слов корпуса “English Borrowings in Russian IT Speech”).



Рис. 1. Визуализация контекстуальных связей ключевых английских заимствований в аудиокорпусе "English Borrowings in Russian IT Speech"

## 2.3. Анализ эффективности системы Yandex SpeechKit в распознавании англицизмов на различных уровнях их освоенности

В рамках нашего исследования также был организован эксперимент, направленный на оценку возможностей современных систем автоматического распознавания речи по распознаванию английских заимствований. Для достижения целей эксперимента использовался API Yandex SpeechKit, так как эта технология широко применяется в Российской Федерации и пользуется большой популярностью в данной сфере.

1. Освоенные англицизмы

Предложение для проверки:

***Маркетинговый*** *отдел разработал впечатляющую кампанию для нового* ***блокбастера****, чтобы привлечь максимальное количество зрителей нового* ***триллера****.*



Рис. 2. Yandex SpeechKit API: распознавание освоенных англицизмов

Вывод:

В результате нашего эксперимента было достигнуто 100% распознавание предложения. Это говорит о высокой эффективности и точности системы распознавания речи, которая успешно справляется с английскими освоенными заимствованиями.

2. Полуосвоенные англицизмы

Предложение для проверки:

*Знаменитая* ***селебрити*** *стала* ***хэдлайнером*** *вечеринки в* ***кокпите*** *самолета, организованной специализированной компанией по* ***хэдхантингу****.*



Рис 3. Yandex SpeechKit API: распознавание полуосвоенных англицизмов

Вывод:
 В результате нашего эксперимента система не смогла распознать слово "кокпит". Остальные слова предложения были успешно распознаны.

**3. Неосвоенные англицизмы**

3.1. Окказионализмы

Предложение для проверки:

*Если он этот* ***таск зафакапит****, а мне опять придется* ***баги фиксить****, я* ***засабмичу*** *шефу репорт и не буду* ***аппрувить энгежмент*** *на следующий проект.*

****

Рис 4. Yandex SpeechKit API: распознавание окказионализмов

Вывод:

В результате нашего эксперимента, при использовании API Yandex SpeechKit было распознано слово "таск", однако остальные слова предложения не были успешно распознаны. Это указывает на то, что следует уделить внимание в нашем исследовании именно окказионализмам.

3.2. Трансплантированные англицизмы

Предложение для проверки:

Ты такой **tough guy**!



Рис 5. Yandex SpeechKit API: распознавание трансплантированных англицизмов

Вывод:

Система распознала трансплантированные англицизмы точно и вывела посредством транслитерации.

 В рамках исследования был проведен эксперимент, целью которого стала оценка способности современных систем автоматического распознавания речи к идентификации английских заимствований разнообразных категорий. В процессе эксперимента использовалась API Yandex SpeechKit, что позволило провести всестороннее тестирование системы на примерах англицизмов, включая полностью ассимилированные, окказиональные и трансплантированные заимствования.

Полученные результаты указывают на то, что системы распознавания речи показывают высокую эффективность при работе с полностью освоенными англицизмами, достигая в некоторых случаях точности в 100%. Однако эффективность заметно снижается при встрече с полуосвоенными и неосвоенными заимствованиями, особенно окказиональными. Например, слова как "кокпит" и другие специализированные термины часто не удавалось распознать, что подчёркивает ограничения текущих технологий в обработке жаргонной и специфической терминологии.

Эти наблюдения подтверждают необходимость дальнейшего развития и оптимизации систем автоматического распознавания речи, чтобы повысить их способность адекватно обрабатывать и идентифицировать разнообразные языковые явления в условиях глобализированного лингвистического контекста. Это особенно критично для корпоративной и IT-сферы, где регулярное использование англицизмов требует от систем не только точности, но и гибкости.

В процессе проведения исследования столкнулись с определёнными ограничениями, связанными с невозможностью дообучения API Yandex SpeechKit. В связи с этим, для достижения научных целей, особенно в части адаптации системы к распознаванию англицизмов, было принято решение использовать технологию Open AI Whisper. Данная система предоставляет возможность дополнительного обучения, что критически важно для реализации задач данного исследования. Процесс дообучения данной системы, а также анализ результатов такого обучения будут тщательно описаны в следующей части исследовательской работы.

## 2.4. Дообучение модели Whisper Open AI на корпусе английских заимствований "English Borrowings in Russian IT Speech"

##

### 2.4.1. Автоматическое распознание английских заимствований без дообучения модели Whisper Open AI

В ходе нашего исследования была оценена способность модели Whisper Open AI распознавать англицизмы в русском IT-жаргоне, причём модель использовалась без предварительного дообучения. Анализ охватил широкий спектр англицизмов, которые активно применяются в профессиональном общении.

Таблица 3. Точность автоматического распознавания англицизмов из корпуса "English Borrowings in Russian IT Speech" моделью Whisper без дообучения

|  |  |  |  |
| --- | --- | --- | --- |
| **Англицизм** | **% Распознавания** | **Англицизм** | **% Распознавания** |
| айтишник | 65,65 | пул реквест | 7,69 |
| алерт | 100,00 | пушить | 81,82 |
| апгрейд | 45,45 | Репортить | 68,57 |
| апрув | 0,00 | сабмит | 12,35 |
| бенчмарк | 31,07 | скоринг | 38,10 |
| бигдата | 33,33 | тимлид | 4,29 |
| бэкап | 8,82 | шарить экран | 66,67 |
| взять\_тайм-аут | 0,00 | шарить | 68,18 |
| ГПУ | 28,57 | юзерфрендли | 13,04 |
| даунгрейд | 6,78 | кешировать | 0,00 |
| дебаг | 42,31 | коммитить | 8,11 |
| девелопер | 28,26 | майнить | 28,41 |
| зафакапить | 68,75 | мерджить | 0,00 |
| зафиксить | 83,33 | мидл | 11,11 |
| пофиксить | 36,36 | Парсить | 75,00 |
| прогать | 0,00 | паттерн | 83,76 |
| профит | 36,84 |  |  |

Мы выявили, что некоторые англицизмы распознавались моделью с высокой точностью, вероятно, благодаря их частому употреблению и ясности контекста. Например:

* "Паттерн" был распознан с успешностью 83.76% (294 из 351 случаев);
* "Пушить" показал успешность в 81.82% случаев (18 из 22 случаев);
* "Парсить" был точно идентифицирован в 75% случаев (48 из 64).

Однако, были и термины, распознавание которых вызвало у модели сложности, особенно если они встречались реже или имели неоднозначное значение. К таким англицизмам относятся:

* "Апрув" и "Прогать", не были распознаны ни в одном из представленных случаев;
* "Тимлид", несмотря на 140 упоминаний, был распознан лишь в 4.29% случаев (6 из 140).

Средний процент точности распознавания приведенных выше англицизмов составляет **35%**. Исходя из этих данных, можно сделать вывод, что модель Whisper демонстрирует значительные возможности для распознавания англицизмов, часто используемых в русском IT-диалекте, даже без дополнительной настройки. Тем не менее, для повышения эффективности распознавания менее распространенных или более сложных для идентификации англицизмов стоит рассмотреть возможность дообучения модели на более разнообразном наборе данных. Это улучшит общую производительность системы и повысит точность распознавания англицизмов в русском языке, что особенно важно для использования в реальных условиях.

### 2.4.2. Описание кода дообучения модели Whisper Open AI

В рамках исследования по дообучению модели Whisper на примере англицизмов в русском IT-жаргоне используется несколько ключевых библиотек Python:

*1. Библиотека Whisper от OpenAI.* Основная библиотека для работы с моделью Whisper.

*2. Transformers.* Часто используется для задач обработки естественного языка, полезна для предобработки текстовых данных и интеграции с моделями.

*3. PyTorch.* Популярная библиотека глубокого обучения, которая может использоваться для управления процессами обучения модели, если необходимы пользовательские слои или функции.

*4. Pandas.* Необходима для работы с данными и анализа, помогает организовать и предобработать набор данных.

*5. NumPy.* Обеспечивает поддержку больших многомерных массивов и матриц, включая широкий спектр высокоуровневых математических функций для работы с этими массивами.

Файл содержит серию ячеек с кодом, предназначенных для выполнения следующих задач:

*1. Загрузка данных.* Код для загрузки и возможной предварительной обработки набора данных, содержащего примеры англицизмов в русской IT-речи.

*2. Подготовка модели. З*агрузка предварительно обученной модели Whisper и настройка ее для дообучения, включая корректировку конфигурации, специфичной для особенностей языковых данных.

*3. Обучение. О*сновная часть кода, где модель Whisper дообучается на подготовленном наборе данных. Это включает в себя определение параметров обучения, возможную корректировку скорости обучения и установку количества эпох.

*4. Оценка. П*осле обучения следует код для оценки производительности модели, проверяя, насколько хорошо дообученная модель справляется с распознаванием англицизмов по сравнению с ее первоначальным состоянием.

*5. Представление результатов. Выводы о производительности.*

**Общая функциональность**

Код предназначен для адаптации модели Whisper к специфическим лингвистическим особенностям англицизмов в русском IT-контексте. Дообучение направлено на повышение точности и эффективности модели в распознавании и обработке этих конкретных видов речи, делая её более устойчивой при работе с реальными данными, для которых она изначально не оптимизирована. Whisper модель включает в себя значительное количество параметров, равное 71.8 миллионам, что свидетельствует о её высокой сложности и потенциале для обработки сложных аудио данных.

Для оценки ошибок и корректировки весов в процессе обучения применяется функция потерь CrossEntropyLoss. Эта функция широко используется в задачах классификации и не вносит дополнительных параметров в модель, поскольку её роль заключается исключительно в расчёте ошибки между предсказанными и истинными значениями.

Анализ параметров модели показал, что из общего числа параметров 52.0 миллиона являются обучаемыми. Эти параметры модифицируются в процессе тренировки для оптимизации производительности модели. Оставшиеся 19.8 миллиона параметров классифицируются как необучаемые, что может включать в себя замороженные веса или другие элементы модели, не изменяющиеся в ходе обучения.

Общий объём памяти, который требуется для хранения параметров модели, составляет приблизительно 287.304 мегабайта. Этот объем важен для учета при развертывании модели в производственной среде, особенно в условиях ограниченных ресурсов вычислительных систем или при работе с большими объемами данных в реальном времени.

Таким образом, исследование структуры и параметров модели Whisper позволяет глубже понять её функциональные возможности и определить оптимальные условия для ее эффективного применения в задачах распознавания речи. Полный код можно найти в Приложении 7. Дообучение модели Whisper Open AI на данных корпуса “English Borrowings in Russian IT Speech”.

### 2.4.3. Процесс дообучения

Для проведения нашего исследования была настроена модель Whisper Open AI с помощью класса конфигурации, который включает в себя ряд параметров, определяющих процесс обучения.

1. Скорость обучения (learning\_rate = 0.0005) – это параметр, который определяет, насколько сильно модель изменяется в ответ на ошибку при каждом обновлении весов. Небольшая скорость обучения помогает обеспечить плавное и точное обновление весов.

2. Вес затухания (weight\_decay = 0.01) – применяется для регуляризации модели, добавляя штраф за слишком большие веса, что помогает избежать переобучения.

3. Epsilon в Adam (adam\_epsilon = 1e-8) – это маленькое число, добавляемое для стабилизации вычислений, предотвращая деление на ноль в алгоритме оптимизации.

4. Шаги подготовки (warmup\_steps = 2) – период, в течение которого скорость обучения постепенно увеличивается до начального значения, что позволяет модели адаптироваться к обучению.

5. Размер пакета (batch\_size = 16) – количество обучающих примеров, обрабатываемых за один шаг. Это важно для управления использованием памяти и скоростью обучения.

6. Количество рабочих процессов (num\_workers = 2) – это количество процессов, используемых для параллельной загрузки данных, что помогает ускорить подготовку и использование данных.

7. Количество эпох обучения (num\_train\_epochs = 10) – полное количество проходов через обучающий набор данных, которое модель должна выполнить в процессе обучения.

8. Шаги накопления градиента (gradient\_accumulation\_steps = 1) – этот параметр позволяет модели накапливать градиенты от нескольких мини-пакетов перед обновлением весов, что может помочь в обучении при ограниченных ресурсах.

9. Частота дискретизации (sample\_rate) – для задач, связанных с аудио, это значение определяет, как часто аудиосигналы преобразуются в цифровую форму для обработки моделью.

Процесс обучения модели включает итерации, в которых данные проходят через модель, оптимизируются её параметры с помощью алгоритма Adam и оценивается ошибка для корректировки весов. Это позволяет модели эффективно учиться, уменьшая ошибки и адаптируясь к задачам, которые перед ней ставились.

### 2.4.4. Оценка и анализ результатов дообучения

В рамках нашего исследования использовалась метрика Character Error Rate (CER) для оценки качества распознавания англицизмов в русском IT-жаргоне с помощью модели Whisper Open AI. Метрика CER, загруженная через функцию `evaluate.load("cer")`, измеряет процент ошибок на уровне символов между предсказанным и истинным текстами. Она вычисляет минимальное количество вставок, удалений и замен символов, необходимых для преобразования предсказанного текста в истинный.

Для вычисления CER были подготовлены два набора данных:

- Истинные тексты (references): это точные расшифровки аудио дорожек, которые служат эталоном для оценки.

- Предсказания модели (predictions): это тексты, которые модель сгенерировала в результате распознавания.

Когда были применены эти данные для вычисления CER, результат показал значение 0.54 Это означает, что около 54% символов в предсказаниях модели оказались ошибочными по сравнению с истинными расшифровками. Такой высокий уровень ошибок указывает на то, что модель испытывает трудности с точным распознаванием англицизмов. Этот результат является сигналом к необходимости дополнительного дообучения модели, а также к возможной корректировке ее настроек для улучшения результатов распознавания. Также необходимо провести еще больше итераций обучения на больших мощностях для улучшения модели.

## 2.5. Выводы к главе II

В рамках исследования был создан и проанализирован аудиокорпус "English Borrowings in Russian IT Speech", что позволило нам глубоко погрузиться в изучение англицизмов в русском IT-языке.

Основной задачей проекта было не просто сформировать корпус англицизмов, но и обеспечить надежную основу для разработки алгоритмов распознавания. Тщательный ручной сбор данных, связанный с IT-тематикой, был проведен путем отбора видеоматериалов на YouTube. Эти материалы обрабатывались, а аудиодорожки расшифровывались в текстовый формат для последующего анализа.

Значительная часть работы была посвящена дообучению модели Open AI Whisper, что включало адаптацию модели под специфику собранных данных с помощью таких инструментов, как Python, библиотеки datasets и transformers.

Кроме того, был проведен статистический и контекстуальный анализ заимствований, что помогло лучше понять, как и в каких контекстах используются англицизмы в русском IT-жаргоне.

Тем не менее, процесс дообучения и оптимизации модели может быть продолжен с использованием описанной в работе методики, что позволит достичь еще более высокой точности в автоматическом распознавании англицизмов. Применяя предложенные подходы, можно адаптировать модель к новым данным и различным языковым контекстам, что обеспечивает ее гибкость и устойчивость. Дополнительно, исследование и внедрение новых алгоритмов и технологий машинного обучения помогут повысить эффективность и точность системы, что сделает ее еще более полезной и надежной в практическом применении.

#

# Заключение

В первой главе исследования были основательно изучены теоретические аспекты автоматического распознавания речи и процесс адаптации английских заимствований в русском языке. Исследование началось с обзора эволюции технологий распознавания речи, от первых механических устройств до современных систем, основанных на принципах глубокого обучения. Особое внимание было уделено сложностям, связанным с многоязычным распознаванием, включая проблемы, возникающие при переключении между языками и особенностями их структур, что имеет ключевое значение при работе с русским и английским языками. Далее тщательно рассматривалось, как англицизмы вливаются в структуру русской речи, и предлагались методы, которые могут оптимизировать их распознавание. Также детально анализировались технические аспекты систем ASR, что позволило глубже понять необходимость их дообучения и адаптации к многоязычным данным.

Во второй главе исследования фокус был сделан на создании и анализе аудиокорпуса "English Borrowings in Russian IT Speech". Процесс включал в себя сбор и систематизацию данных, а также глубокое дообучение модели Open AI Whisper с применением разнообразных программных инструментов, улучшающих её способность к распознаванию и интерпретации заимствований. Статистический и контекстуальный анализы помогли нам лучше осмыслить употребление заимствований в различных контекстах, что послужило основой для дальнейшего усовершенствования процесса распознавания.

Итоги обеих глав подтверждают важность интегрированного подхода к обучению и адаптации систем автоматического распознавания речи. Данные результаты исследования являются критически важными для развития эффективных многоязычных систем ASR, которые могут адекватно функционировать в условиях языкового разнообразия и динамично меняющихся языковых интеракций современного мира. Исследование создаёт основу для последующих разработок в этой области, способствуя улучшению интерфейсов взаимодействия человека и машины в мировом информационном пространстве.

На основе полученных данных можно предложить рекомендации для дальнейших исследований и практического применения. Важно провести углубленное изучение случаев, когда говорящий артикулирует английские заимствования с нормативным и ненормативным произношением, что поможет улучшить адаптивность и точность систем автоматического распознавания речи.

 Также необходимо разработать и тестировать новые алгоритмы и методики дообучения, особенно для обработки код-свитчинга и заимствованных фраз. Рекомендуется стратегическое сотрудничество с лингвистами и IT-специалистами для разработки более точных и универсальных ASR систем, а также проведение масштабных испытаний для оценки производительности и устойчивости систем в реальных условиях.

Перспективы применения результатов особенно значимы в корпоративной и IT-среде. Они могут значительно улучшить коммуникацию в рабочих коллективах, а также быть использованы для разработки многоязычных платформ, которые повышают эффективность взаимодействия и удобство работы в международных компаниях.

Таким образом, результаты исследования могут стать основой для дальнейших улучшений в области машинного обучения и лингвистики, способствуя созданию более естественных и интеллектуальных систем человеко-машинного взаимодействия.

# Список литературы

1. Валуйцева И.И., Филатов И.Е. ПОДХОДЫ К ЯЗЫКОВОМУ И АКУСТИЧЕСКОМУ МОДЕЛИРОВАНИЮ В РАСПОЗНАВАНИИ РЕЧИ. Вопросы прикладной лингвистики. 2019. № 4 (36). С. 7-31. URL: https://www.elibrary.ru/item.asp?id=42799712 (дата обращения: 14.05.2023).
2. Дьяков, А.И., Золотарева, Е.А. Адаптационная модель англицизмов: сборник трудов конференции // Научные исследования: от теории к практике : материалы IX Междунар. науч.-практ. конф. (Чебоксары, 10 июля 2016 г.) / редкол.: О.Н. Широков [и др.] – Чебоксары: Центр научного сотрудничества «Интерактив плюс», 2016. – С. 245-255. – ISSN 2413-3957. – DOI 10.21661/r-112194.
3. Лингвистический энциклопедический словарь / гл. ред. В.Н. Ярцева. – 2-е изд., доп. – М. : Большая рос. энцикл., 2002. – 709 с. URL: <https://tapemark.narod.ru/les/> (дата обращения: 10.05.2024).
4. Николова Д., (2015). Коммуникация в сфере информационных технологий и жаргон как одно из ее средств // Служение слову. № 3. URL: https://www.shu.bg/wp-content/uploads/file-manager-advanced/users/faculties/fhn/izdaniya/sluzhenie-slovu/Sluzhenie-2015.pdf#page=111 (дата обращения: 14.05.2023).
5. A Brief History of ASR: Automatic Speech Recognition. URL: [https://medium.com/descript/a-brief-history-of-asr-automatic-speech-recognition-b8f338d4c0e5] (дата обращения: 23.09.2023).
6. Arora, A., Chang, C.-C., Rekabdar, B., Povey, D., Etter, D., Raj, D., Hadian, H., Trmal, J., García-Perera, L.P., Watanabe, S., Manohar, V., Shao, Y., & Khudanpur, S. Using ASR Methods for OCR // ResearchGate. 2019. URL: https://www.researchgate.net/publication/339027380\_Using\_ASR\_Methods\_for\_OCR (дата обращения: 30.11.23).
7. arXiv: Recent Advances in Automatic Speech Recognition. URL: [https://arxiv.org/abs/2108.00084] (дата обращения: 12.10.2023).
8. Baktash, J., & Dawodi, M. (2023). GPT-4: A Review on Advancements and Opportunities in Natural Language Processing. ArXiv. URL: https://arxiv.org/abs/2305.03195 (дата обращения: 03 .04.24).
9. Bellegarda, J. (2000). Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE, 88, 1279-1296. URL: [https://doi.org/10.1109/5.880084] (дата обращения: 15.10.2023).
10. Bhatt, S., Jain, A., & Dev, A. (2020). Acoustic Modeling in Speech Recognition: A Systematic Review. International Journal of Advanced Computer Science and Applications, 11. (дата обращения: 30.09.2023).
11. Cai, J., & Liu, Y. (2018). Research on English pronunciation training based on intelligent speech recognition. International Journal of Speech Technology, 21, 633-640. URL: https://doi.org/10.1007/s10772-018-9523-8 (дата обращения: 17.10.2023).
12. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. arXiv preprint arXiv:2006.13979. URL: https://arxiv.org/abs/2006.13979 (дата обращения: 14.10.2023).
13. Crowson, M., Lee, J., Hamour, A., Mahmood, R., Babier, A., Lin, V., Tucci, D., & Chan, T. (2020). AutoAudio: Deep Learning for Automatic Audiogram Interpretation. Journal of Medical Systems, 44, 1-7. URL: https://doi.org/10.1007/s10916-020-01627-1 (дата обращения: 28.09.2023).
14. Deepgram: The History of Speech Recognition Technology. URL: [https://deepgram.com/learn/the-history-of-automatic-speech-recognition] (дата обращения: 28.09.2023).
15. Dighe, P., Luyet, G., Asaei, A., & Bourlard, H. (2016). Exploiting low-dimensional structures to enhance DNN based acoustic modeling in speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5690-5694. URL: [https://doi.org/10.1109/ICASSP.2016.7472767] (дата обращения: 25.09.2023).
16. Dokuz, Y., & Tufekci, Z. (2021). Mini-batch sample selection strategies for deep learning based speech recognition. Applied Acoustics, 171, 107573. URL: [https://doi.org/10.1016/j.apacoust.2020.107573] (дата обращения: 18.10.2023).
17. Haeb-Umbach, R., Heymann, J., Drude, L., Watanabe, S., Delcroix, M., University, T., Germany, H., Aachen, A., University, J., , B., , U., Laboratories, N., , K., & , J. (2020). Far-Field Automatic Speech Recognition. Proceedings of the IEEE, 109, 124-148. URL: https://doi.org/10.1109/JPROC.2020.3018668 (дата обращения: 28.09.2023).
18. Hu, Z., Lan, Y., Wang, L., Xu, W., Lim, E.-P., Lee, R., Bing, L., & Poria, S. (2023). LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. ArXiv. URL: https://arxiv.org/abs/2304.01933 (дата обращения: 10.04.24).
19. Huang, Z., Siniscalchi, S., & Lee, C. (2016). A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition. Neurocomputing, 218, 448-459. URL: https://doi.org/10.1016/j.neucom.2016.09.018 (дата обращения: 30.09.2023).
20. IBM: What is Automatic Speech Recognition? URL: [https://www.ibm.com/cloud/learn/speech-recognition] (дата обращения: 05.11.2023).
21. Jain, A., Upreti, M., & Jyothi, P. Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning // CSE IIT Bombay. 2018. URL: https://www.cse.iitb.ac.in/~pjyothi/files/IS18b.pdf (дата обращения: 22.10.23).
22. Jermsittiparsert, K., Abdurrahman, A., Siriattakul, P., Sundeeva, L., Hashim, W., Rahim, R., & Maseleno, A. (2020). Pattern recognition and features selection for speech emotion recognition model using deep learning. International Journal of Speech Technology, 23, 799 - 806. URL: https://www.researchgate.net/publication/344269353\_Pattern\_recognition\_and\_features\_selection\_for\_speech\_emotion\_recognition\_model\_using\_deep\_learning (дата обращения: 17.10.2023).
23. Joy, N. M., Umesh, S. Improving Acoustic Models in TORGO Dysarthric Speech Database // IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2017. URL:[https://ieeexplore.ieee.org/document/8343813](https://ieeexplore.ieee.org/document/8343813) (дата обращения:18.10.2023).
24. Kipyatkova Irina. 2022. Investigation of Transfer Learning for End-to-End Russian Speech Recognition. In Speech and Computer: 24th International Conference, SPECOM 2022, Gurugram, India, November 14–16, 2022, Proceedings. Springer-Verlag, Berlin, Heidelberg, 349–357. https://doi.org/10.1007/978-3-031-20980-2\_30
25. Liu, H., Teng, Z., Cui, L., Zhang, C., Zhou, Q., & Zhang, Y. (2023). LogiCoT: Logical Chain-of-Thought Instruction-Tuning Data Collection with GPT-4. ArXiv. URL: https://arxiv.org/abs/2305.12147 (дата обращения: 17.03.24).
26. Marcus, G. (2018). Deep Learning: A Critical Appraisal. ArXiv, abs/1801.00631. URL: https://arxiv.org/abs/1801.00631 (дата обращения: 28.09.2023).
27. Nejedly, P., Kremen, V., Sladky, V., Nasseri, M., Guragain, H., Klimeš, P., Cimbalnik, J., Varatharajah, Y., Brinkmann, B., & Worrell, G. (2019). Deep-learning for seizure forecasting in canines with epilepsy. Journal of Neural Engineering, 16. URL: https://doi.org/10.1088/1741-2552/ab172d (дата обращения: 17.10.2023).
28. Noyes, J. M., Haigh, R., & Starr, A. F. (1989). Automatic speech recognition for disabled people. Applied Ergonomics, 20(4), 293–298. URL: https://doi.org/10.1016/0003-6870(89)90193-2 (дата обращения: 30.09.2023).
29. NVIDIA: Automatic Speech Recognition (ASR) Technology. URL: [https://developer.nvidia.com/blog/an-easy-introduction-to-speech-ai/] (дата обращения: 21.11.2023).
30. Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction Tuning with GPT-4. ArXiv. URL: https://arxiv.org/abs/2304.03277 (дата обращения: 01.04.24).
31. Putland, R., Ranjard, L., Constantine, R., & Radford, C. (2018). A hidden Markov model approach to indicate Bryde’s whale acoustics. Ecological Indicators, 84, 479-487. URL: [https://doi.org/10.1016/J.ECOLIND.2017.09.025] (дата обращения: 07.11.2023).
32. Qiao, Y., & Minematsu, N. (2010). A Study on Invariance of $f$-Divergence and Its Application to Speech Recognition. IEEE Transactions on Signal Processing, 58, 3884-3890. URL: [https://doi.org/10.1109/TSP.2010.2047340] (дата обращения: 18.10.2023).
33. Rasipuram, R., & Magimai.-Doss, M. (2015). Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model. Speech Commun., 68, 23-40. URL: [https://doi.org/10.1016/j.specom.2014.12.006] (дата обращения: 05.11.2023).
34. Rodero, E. (2016). Influence of speech rate and information density on recognition: The moderate dynamic mechanism. Media Psychology, 19(2), 224–242. URL: https://repositori.upf.edu/bitstream/handle/10230/35734/Rodero\_med\_infl.pdf?sequence=1&isAllowed=y (дата обращения: 25.09.2023).
35. Schultz, B.G., Tarigoppula, V.S.A., Noffs, G. et al. (2021). Automatic speech recognition in neurodegenerative disease. Int J Speech Technol 24, 771–779. URL: https://doi.org/10.1007/s10772-021-09836-w (дата обращения: 30.09.2023).
36. Sennrich, R. (2015). Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation. Transactions of the Association for Computational Linguistics, 3, 169-182. URL: [https://doi.org/10.1162/tacl\_a\_00131] (дата обращения: 23.09.2023).
37. Takahashi, S., & Tanaka-Ishii, K. (2019). Evaluating Computational Language Models with Scaling Properties of Natural Language. Computational Linguistics, Just Accepted, 1-32. URL: [https://doi.org/10.1162/COLI\_a\_00355] (дата обращения: 17.10.2023).
38. The History of Voice Recognition Technology. URL: [https://www.condecosoftware.com/blog/a-history-of-voice-recognition-technology](дата обращения: 17.10.2023).
39. The Jomr's Guide To Automatic Speech Recognition System. URL: [https://www.condecosoftware.com/blog/a-history-of-voice-recognition-technology](дата обращения: 05.11.2023).
40. Thompson, N. C., Greenewald, K. H., Lee, K., Manso, G. F. The Computational Limits of Deep Learning // ArXiv. 2020. URL: [https://arxiv.org/abs/2007.05558](https://arxiv.org/abs/2007.05558) (дата обращения: 18.10.2023).
41. Waibel, A., Soltau, H., Schultz, T., Schaaf, T., & Metze, F. Multilingual Speech Recognition // ResearchGate. 1997. URL: https://www.researchgate.net/publication/2897682\_Multilingual\_Speech\_Recognition (дата обращения: 15.09.23).
42. Wilpon Jay G. Improving speaker consistency in an automatic speech recognition framework Author links open overlay panelLinda A. Roberts, Jay G. Wilpon, Dennis E. Egan, Jean Bakk. URL: https://doi.org/10.1016/S0885-2308(86)80011-0 (дата обращения: 15.10.2023).
43. Zolotareva, Elena. (2016). Adaptation model of anglicisms. С. 245-255. DOI 10.21661/r-112194. URL: https://www.researchgate.net/publication/308090179\_Adaptation\_model\_of\_anglicisms (дата обращения: 02.05.2023).
44. Zoom’s new AI features help you catch up on meetings you’re late to. URL: [https://www.theverge.com/2023/3/27/23658047/zoom-ai-features-meeting-summaries-mail-calendar] (дата обращения: 17.10.2023).
45. Zoom's new AI tools will soon summarize your meetings for you. URL: [https://www.zdnet.com/article/zooms-new-ai-tools-will-soon-summarize-your-meetings-for-you/] (дата обращения: 23.09.2023).

# Приложения

**Приложение 1. Корпус аудиоданных “English Borrowings in Russian IT Speech”**

1. [Google Drive Link](https://drive.google.com/drive/folders/1WdUelLqig_dZvFASiux4b3iI7H1A-AU5?usp=sharing)

**Приложение 2. Список ключевых слов, использованных для поиска видеоматериалов**

|  |  |
| --- | --- |
| **Keyword** | **Keyword** |
| Искусственный интеллект (AI) | Metaverse |
| Машинное обучение (ML) | NFT (Non-fungible tokens) |
| Глубокое обучение (Deep Learning) | Web Development |
| Блокчейн | Mobile Development |
| Облачные технологии | Augmented Reality Development |
| Кибербезопасность | Virtual Reality Development |
| Big Data | Cyber-Physical Systems |
| Интернет вещей (IoT) | Sustainable Tech |
| Самоуправляемые автомобили | Green Computing |
| Виртуальная реальность (VR) | E-Waste Recycling |
| Дополненная реальность (AR) | Privacy Tech |
| Квантовые вычисления | GDPR |
| 5G технологии | Ethical Hacking |
| Робототехника | Digital Forensics |
| Автоматизация | Cloud Security |
| DevOps | SaaS (Software as a Service) |
| Контейнеризация | PaaS (Platform as a Service) |
| Microservices (Микросервисы) | IaaS (Infrastructure as a Service) |
| Финтех | Serverless Computing |
| Смарт-контракты | Blockchain as a Service |
| Python | Quantum Cryptography |
| JavaScript | Deepfakes |
| Java | Edge AI |
| C# | Generative Adversarial Networks (GANs) |
| PHP | Swift |
| Ruby | Kotlin |
| Microsoft Azure | Смарт-дома |
| Google Cloud Platform | Электронная коммерция |
| DevSecOps | UI/UX дизайн |
| Agile методологии | Системы распознавания образов |
| Scrum | Chatbots |
| Канбан (Kanban) | Natural Language Processing (NLP) |
| CI/CD (Continuous Integration/Continuous Deployment) | Computer Vision |
| RESTful API | Системы управления базами данных (СУБД) |
| GraphQL | Системы контроля версий |
| Git | Blockchain Platforms |
| Linux | Smart Cities |
| Unix | SQL |
| Windows Server | NoSQL |
| Безопасность сетей | Docker |
| Penetration тесты | Kubernetes |
| Криптография | TensorFlow |
| Data Science | PyTorch |
| Data Analytics | Scikit-learn |
| Data Mining | Pandas |
| Data Visualization | Numpy |
| Tableau | Cloud Computing |
| Power BI | Amazon Web Services (AWS) |
| Business Intelligence | Microsoft Azure |
| Edge Computing | Swift |
| Digital Twin | Kotlin |
| Электронная коммерция | Смарт-дома |
| UI/UX дизайн | Web Development |
| Mobile Development |  |

**Приложение 3. Экстрактор аудио по временным рамкам из YouTube видео корпуса “English Borrowings in Russian IT Speech”**

1. [Google Drive Link](https://drive.google.com/file/d/16c9DCYzp5JsYWPuj468BEorJUeXqkr2I/view?usp=drive_link)

**Приложение 4. Уникальные слова корпуса “English Borrowings in Russian IT Speech”**

1. [Google Drive Link](https://drive.google.com/drive/folders/1Zmxlm7VaF9i8ERcw9FY3RSqovHJgjix9?usp=drive_link)

**Приложение 5. Контекст ключевых слов корпуса “English Borrowings in Russian IT Speech”**

1. [Google Drive Link](https://drive.google.com/drive/folders/16FJBB1CYQQXYwXvdjiJVnbZc3EO2BtFN?usp=drive_link)

**Приложении 6. Дообучение модели Whisper Open AI на данных корпуса “English Borrowings in Russian IT Speech”.**

1. [Google Colab Link](https://colab.research.google.com/drive/16Xjy-z9FsYsEozzPSI7hbO2OioipqdW9?usp=sharing)