МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ

Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

Разработка интеллектуального клиент-серверного сервиса распознавания речи и эмоционального состояния пользователя на основе методов машинного обучения

АВТОРЕФЕРАТ

студента 4 курса, 441 группы

направления 02.03.03 «Математическое обеспечение и администрирование информационных систем»

факультета Компьютерных наук и информационных технологий

Бережнова Артёма Валерьевича

Научный руководитель
к.э.н., доцент _____ Кабанова Л. В.
(подпись, дата)

Зав. кафедрой
к.ф.-м.н., доцент _____ Огнева М. В.
(подпись, дата)

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы исследования. В современную цифровую эпоху голосовые технологии стали неотъемлемой частью экосистем умных домов и мобильных устройств. Однако большинство существующих систем (таких как Siri, Алиса, Google Assistant) обладают существенным ограничением: они анализируют преимущественно семантическое содержание речи (текст), игнорируя её эмоциональную и интонационную составляющую. Согласно исследованиям в области психологии коммуникаций, вербальная информация составляет лишь малую часть передаваемого смысла, в то время как акустические параметры голоса несут критически важные данные о состоянии говорящего. Разработка систем, способных распознавать эмоции в режиме реального времени (Speech Emotion Recognition, SER), является необходимым шагом для создания по-настоящему антропоцентричного и эмпатичного искусственного интеллекта. Данная работа посвящена разработке программного комплекса, который ложится в основу интеллектуального ассистента Stillara, создаваемого в рамках программы «Стартап как диплом».

Объектом исследования являются процессы цифровой обработки аудиосигналов и автоматизированного анализа эмоционального состояния пользователя на основе голосовых данных.

Предметом исследования выступают алгоритмы машинного обучения, методы извлечения акустических признаков речевого сигнала, а также архитектурные паттерны построения клиент-серверных систем для интеграции модулей распознавания речи и определения эмоций.

Целью выпускной квалификационной работы является проектирование и программная реализация интеллектуальной клиент-серверной системы распознавания речи и эмоционального состояния пользователя (Backend и AI-модули), обеспечивающей высокую точность классификации и возможность бесшовной интеграции в интерфейс приложения.

Задачи исследования:

Провести анализ современных методов цифровой обработки сигналов и существующих подходов к распознаванию эмоций (SER).

Спроектировать клиент-серверную архитектуру системы и схему реляционной базы данных для хранения мультимедийных данных и метаданных.

Реализовать программный модуль извлечения акустических признаков (временных, спектральных и кепстральных характеристик) с использованием специализированных библиотек.

Разработать и обучить классификатор эмоциональных состояний на базе алгоритмов машинного обучения, провести сравнительный анализ их эффективности.

Реализовать Backend-часть системы (REST API) и Frontend-модуль для взаимодействия с пользователем и внешними сервисами.

Провести комплексное тестирование, оценить точность распознавания и проверить стабильность системы под нагрузкой.

Новизна работы заключается в следующем:

Сформирован и аугментирован сбалансированный русскоязычный датасет для распознавания эмоций, объединяющий данные публичного корпуса RAVDESS и собственную фонотеку, что позволило повысить релевантность модели для русскоязычной среды.

Экспериментально обоснован выбор ансамблевого метода Extra Trees для классификации эмоций по 40 акустическим признакам, который продемонстрировал оптимальный баланс между смещением и дисперсией (accuracy 70%) на выборках среднего объема, превосходя классические методы градиентного бустинга и линейные модели.

Предложен механизм автоматической проверки совместимости размерности признаков при загрузке ML-модели, обеспечивающий обратную

совместимость и отказоустойчивость backend-сервиса при изменении пайплайна извлечения фич.

Практическая значимость работы состоит в создании готового к внедрению программного продукта (интеллектуального ассистента Stillara), который учитывает эмоциональный контекст пользователя. Разработанная архитектура позволяет масштабировать систему, а полученные в ходе эксперимента выводы о важности акустических признаков (MFCC, RMS-энергия, спектральный центроид) могут быть использованы в дальнейших исследованиях в области аффективных вычислений.

Положения, выносимые на защиту:

Архитектура клиент-серверного сервиса, включающая асинхронный Backend на FastAPI, реактивный Frontend на Vue.js 3 и локальный ML-инференс, обеспечивающая задержку обработки менее 2 секунд.

Методика извлечения 40 акустических признаков включает нормализацию и шумоподавление, после чего формируется вектор для классификации. MFCC (13 + дельты) моделируют восприятие спектра ухом, отражая тембр и фонемы. Chroma (12) описывают гармоническую структуру через 12 полутонов. Spectral Contrast (7) показывает разницу между пиками и впадинами спектра, характеризуя яркость звука. ZCR (1) — частота пересечения нуля, отражает зашумлённость и атаку. RMS (1) — энергия сигнала, показатель громкости. Суммарно 34 признака, с дельта-коэффициентами MFCC — 40.

Результаты экспериментального исследования 17 алгоритмов машинного обучения, доказывающие превосходство ансамблевых методов (Extra Trees) над линейными и байесовскими моделями в задачах SER из-за нелинейной природы акустических данных и мультиколлинеарности признаков.

Комплекс мер информационной безопасности, включающий JWT-аутентификацию, httpOnly-куки, шифрование биометрических данных (AES-256-GCM) и защиту от DoS-атак.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе, посвященном теоретическим основам разработки интеллектуального сервиса, проведен обзор существующих систем, таких как Google Speech Recognition, Microsoft Azure, OpenSMILE и Beyond Verbal, выявлены их преимущества и недостатки. Показано, что коммерческие решения часто закрыты для кастомизации, а академические, например OpenSMILE, требуют сложной интеграции. Обоснован выбор клиент-серверной архитектуры, поскольку тяжелые вычислительные задачи, связанные с инференсом ML-моделей, консолидируются на сервере, что позволяет использовать GPU- или CPU-оптимизированное оборудование и централизованно обновлять модели.

Клиентская часть при этом отвечает за захват аудио через WebRTC API, предобработку, включая нормализацию амплитуды и ресемплинг до 16 кГц, и передачу данных по протоколу HTTPS. Подробно описан математический аппарат обработки аудиосигналов: поскольку речевой сигнал является нестационарным, применяется процедура фрейминга с окнами по 20–40 миллисекунд и наложением 50% с использованием окна Хэмминга для минимизации утечки спектра. Ключевым этапом является извлечение Mel-кепстральных коэффициентов, имитирующих нелинейное восприятие частоты человеческим слухом, а дополнительно извлекаются спектральные признаки, включая центроид, ширину спектра, roll-off и контраст, а также временные характеристики, такие как ZCR, RMS-энергия и основная частота F0.

Рассмотрены методы машинного обучения для многоклассовой классификации, включая SVM, Random Forest, Gradient Boosting, CNN и RNN/LSTM, и обоснована необходимость стратифицированного разделения выборки и использования метрик Precision, Recall, F1-score и Macro-average для оценки моделей в условиях возможного дисбаланса классов. Также описаны современные подходы к фронтенд-разработке на Vue.js 3 с компонентным подходом, реактивностью и Pinia для управления состоянием, а также

принципы безопасности веб-приложений, включая хэширование паролей с «солью», JWT, защиту от XSS и CSRF, а также Rate Limiting.

Во втором разделе, посвященном разработке и исследованию системы, описывается проектирование архитектуры и базы данных. Архитектура формализована с помощью UML-диаграмм и включает клиентское SPA на Vue.js, сервер на FastAPI, модуль ASR на базе Google Speech API, модуль SER с локальной моделью Extra Trees и СУБД PostgreSQL. Спроектирована схема базы данных в третьей нормальной форме: таблица audio_sessions хранит метаданные записей, при этом сами аудиофайлы хранятся в файловой системе, а в базе данных — только путь к ним; таблица analysis_results со связью один-к-одному хранит распознанный текст, детектированную эмоцию, уверенность модели и полный вектор вероятностей в формате JSONB для последующей аналитики. Применено партиционирование и составные индексы для ускорения выборок. Далее описывается разработка модуля обработки аудиосигналов: реализован класс AudioFeatureExtractor набазе библиотеки LibROSA, извлекающий сорок признаков, включая тринадцать MFCC, двенадцать хроматических признаков, семь показателей спектрального контраста, а также ZCR, RMS, спектральный центроид, rolloff, полосу пропускания, темп, а также гармоническую и перкуссионную компоненты. Для достижения задержки менее двух секунд применена многопоточная обработка: распознавание речи через сетевой вызов к Google API и классификация эмоций с локальными вычислениями запускаются параллельно в разных потоках, что позволило сократить медианное время обработки с 1120 до 890 миллисекунд. Особое внимание уделено обучению и сравнительному анализу моделей: сформирован датасет из 1880 записей по пяти классам — радость, грусть, злость, непонимание и беспокойство, при этом применена аугментация путем добавления шума и сдвига темпа, увеличившая выборку в полтора раза. Проведен масштабный эксперимент по сравнению семнадцати

алгоритмов машинного обучения, результаты которого показали, что лучший результат демонстрирует Extra Trees с точностью 70,0% и F1-мерой 0,698, второе место занимает Random Forest с точностью 66,6%, тогда как линейные модели, такие как логистическая регрессия и LDA, показали низкую точность в пределах 50–57% из-за нелинейной природы разделения эмоций. Наивный байесовский классификатор оказался неприменим с точностью 44,9% из-за нарушения предположения о независимости признаков, поскольку мультиколлинеарность MFCC-коэффициентов достигает 0,8–0,9, а нейронные сети в виде MLP не оправдали ожиданий с точностью 48,7% из-за недостаточного объема данных для обучения более восьми тысяч параметров.

Анализ важности признаков выявил, что ключевыми для классификации являются mfcc_1, отражающий спектральную огибающую, chroma_9, связанный с тональностью, и rms_energy, характеризующий громкость, при этом основные ошибки модели связаны с акустической схожестью классов «беспокойство» и «непонимание». В рамках разработки Backend и Frontend частей Backend реализован на FastAPI с эндпоинтом POST /api/process-voice, который принимает аудио в формате WebM, конвертирует его в WAV через FFmpeg, параллельно запускает ASR и SER и возвращает JSON-ответ. Реализован механизм автоматического создания демо-модели на синтетических данных, если файл обученной модели отсутствует или несовместима размерность признаков. Frontend реализован на Vue.js 3: компонент ChatPage.vue использует WebRTC и MediaRecorder для захвата голоса, реализован интуитивный пользовательский интерфейс с кнопкой «зажми для записи», анимацией пульсации и индикатором «ИИ анализирует...». После ответа сервера бот генерирует эмпатичный ответ с задержкой 1,2 секунды, учитывая распознанную эмоцию, также разработан дашборд с SVG-графиками, включая радарную диаграмму и столбчатую диаграмму, для отслеживания динамики настроения пользователя. Проведено End-to-End тестирование с помощью Cypress и

нагрузочное тестирование с использованием Locust: при нагрузке в 50 пользователей среднее время ответа составило 1,2 секунды, а 95-й перцентиль — 2 секунды. Для деплоя на домашний сервер без статического IP использован Cloudflare Tunnel, что обеспечило автоматическое получение SSL-сертификатов, DDoS-защиту и скрытие реальной инфраструктуры. В экспериментальном исследовании методов машинного обучения представлен детальный математический и физический анализ результатов, доказывающий, что ансамблевые методы на основе бэггинга превосходят бустинг на зашумленных аудиоданных среднего объема, поскольку усреднение множества деревьев снижает дисперсию и повышает робастность к индивидуальным особенностям тембра дикторов. Сделан вывод о необходимости перехода к сверточным нейронным сетям на спектрограммах или рекуррентным сетям типа LSTM при увеличении датасета до более пяти тысяч записей для преодоления порога точности в 80%.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы была успешно спроектирована, разработана и исследована интеллектуальная клиент-серверная система распознавания речи и определения эмоций по голосу для приложения-помощника Stillara.

Основные результаты и выводы работы:

Разработана модульная архитектура сервиса, разделяющая ответственность между клиентским приложением (Vue.js), серверной логикой (FastAPI) и AI-модулями. Интеграция с Google Speech API и локальным ML-классификатором позволяет обеспечивать высокую скорость отклика.

Реализован конвейер цифровой обработки аудиосигналов на базе LibROSA, извлекающий 40 информативных акустических признаков (MFCC, хроматические, спектральные и временные характеристики).

Сформирован сбалансированный датасет из 1880 записей на русском и

английском языках. Проведенное сравнительное исследование 17 алгоритмов машинного обучения доказало, что для задач SER на датасетах среднего размера оптимальным является алгоритм Extra Trees, достигший точности 70.0% и F1-меры 0.698.

Выявлены физические и математические ограничения линейных и байесовских моделей в данной предметной области, обусловленные нелинейностью акустических паттернов и высокой мультиколлинеарностью спектральных признаков.

Создан полнофункциональный пользовательский интерфейс, включающий голосовой чат с визуализацией эмоций, дашборд отслеживания настроения и форум поддержки с системой анонимных публикаций.

Реализован многоуровневый механизм безопасности, соответствующий требованиям 152-ФЗ и GDPR, включающий шифрование биометрических данных, защиту от XSS/CSRF и Rate Limiting.

Система успешно протестирована и развернута в среде, приближенной к продакшн, с использованием Cloudflare Tunnel для безопасного доступа из публичной сети.

Практическая значимость работы заключается в создании основы для эмпатичных AI-систем, способных адаптировать тон и содержание ответа в зависимости от психоэмоционального состояния пользователя, что открывает новые перспективы в области разработки цифровых ассистентов, систем телемедицины и клиентской поддержки.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Mozilla Developer Network. WebRTC API [Электронный ресурс]. – URL: https://developer.mozilla.org/en-US/docs/Web/API/WebRTC_API (дата обращения: 15.09.2025).

2. Google Cloud Speech-to-Text Documentation [Электронный ресурс] // Google Cloud. – URL: https://cloud.google.com/speech-to-text (дата обращения: 23.10.2025)

3. McFee B., Raffel C., Liang D. et al. librosa: Audio and Music Signal Analysis in Python // Proceedings of the 14th Python in Science Conference. – 2015. – P. 18–25. (дата обращения: 30.10.2025)

4. Fielding R. Architectural Styles and the Design of Network-based Software Architectures. – University of California, Irvine, 2000. – 180 p. (дата обращения: 19.09.2025)

5. Richardson L., Amundsen M. RESTful Web APIs. – O'Reilly Media, 2013. – 408 p. (дата обращения: 19.09.2025)

6. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186. (дата обращения: 29.09.2025)

7. Szegedy C., Liu W., Jia Y. et al. Going Deeper with Convolutions // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 1–9. (дата обращения: 05.10.2025)