МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра математической теории упругости и биомеханики

Проектирование системы поддержки преодоления языкового барьера

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 442 группы

направления 09.03.03 – Прикладная информатика

механико-математического факультета

Шишовой Алёны Сергеевны

Научный руководитель
доцент, к.ю.н. Р.В. Амелин

Зав. кафедрой
зав. кафедрой, д.ф.-м.н., профессор Л.Ю. Коссович

Саратов 2026

Введение. Актуальность темы. Русский жестовый язык (РЖЯ) является самостоятельной лингвистической системой и официальным средством коммуникации для глухих и слабослышащих граждан Российской Федерации. Несмотря на это, сохраняется устойчивый коммуникационный барьер между слышащим и жестовым сообществами. Современные системы автоматического распознавания жестовых языков преимущественно ориентированы на англоязычные датасеты, требуют облачной инфраструктуры и мощных графических ускорителей, а также часто игнорируют лингвистическую специфику РЖЯ. В условиях ужесточения требований к защите биометрических данных и необходимости обеспечения работы в среде с нестабильным интернет-соединением возрастает потребность в легковесных локальных решениях (on-device), оптимизированных для стандартных процессоров и прошедших строгую валидацию на уровне пользовательской выборки.

Целью работы является адаптация и экспериментальная оценка прототипа локальной системы двунаправленного перевода между РЖЯ и русским языком с архитектурной асимметрией: основное внимание уделено распознаванию динамических жестов (направление «жесты — текст») с использованием последовательностных моделей, оптимизированных для работы на CPU без облачных зависимостей.

Для достижения цели решаются следующие **задачи**:

- Провести анализ существующих технических решений в области автоматического распознавания жестовых языков и выявить их ограничения применительно к РЖЯ.
- Сформулировать функциональные и нефункциональные требования к системе, обосновать выбор технологического стека и архитектуры локального развёртывания.
- Сформировать сбалансированное подмножество открытого датасета Slovo (30–50 глосс), разработать и автоматизировать конвейер предобработки: извлечение скелетных признаков, нормализацию координат, сглаживание траекторий и выравнивание длины последовательностей.
- Адаптировать и дообучить архитектуру BiLSTM для классификации динамических жестов, применив стратегию разделения выборки по

идентификаторам участников (user-based split), исключающую пересечение пользователей между обучающей и тестовой выборкой.
- Оптимизировать инференс модели для работы на CPU путём экспорта в формат TensorFlow Lite.
- Провести экспериментальную оценку качества распознавания, измерить производительность системы, сформулировать ограничения и перспективы её развития.

В работе используется открытый видеодатасет Slovo (Sber AI), из которого выделено сбалансированное подмножество из 42 частотных бытовых и учебных глосс. Общий объём отобранных записей после фильтрации и балансировки составил 970 видеороликов. Для извлечения признаков применяется фреймворк MediaPipe Hands, для построения и обучения модели – библиотека TensorFlow/Keras, для оптимизации и развёртывания – TensorFlow Lite и фреймворк Streamlit.

Структура работы. Выпускная квалификационная работа состоит из введения, трёх глав, заключения и списка использованных источников.
- Глава 1 "Анализ предметной области, существующих решений и требований к системе"посвящена аналитическому обзору предметной области и постановке задачи.
- Глава 2 "Проектирование и реализация системы распознавания жестов"описывает проектирование и реализацию системы распознавания жестов.
- Глава 3 "Реализация и тестирование прототипа"содержит результаты экспериментальной оценки, анализ производительности и демонстрацию работоспособности прототипа.

В первой главе проведён анализ лингвистических и технических особенностей РЖЯ как объекта компьютерного зрения. Установлено, что РЖЯ опирается на одновременную комбинацию пространственных, кинематических и не-мануальных параметров (мимика, движение корпуса). Игнорирование не-мануальных маркеров в базовых моделях признаётся системным ограничением, требующим явного указания в описании прототипа. С технической точки зрения задача относится к последовательностной классификации пространственно-временных рядов и сопряжена с временной вариативностью

исполнения, пространственной неоднозначностью и биометрической чувствительностью видеопотока.

Проанализированы зарубежные облачные сервисы и академические прототипы на базе 3D CNN, трансформеров и графовых свёрток. Выявлено, что большинство систем требуют специализированного оборудования, GPU-ускорения и стабильного интернет-соединения, что ограничивает их применимость в условиях локального развёртывания и нарушает принципы конфиденциальности биометрических данных. В российской практике ключевым достижением является публикация датасета Slovo, однако публичные реализации часто не включают полный цикл локального деплоя и детерминированной предобработки.

На основе проведённого анализа, были сформулированы следующие требования к системе. Таким образом, функциональные требования включают приём видеопотока, извлечение скелетных признаков, классификацию в один из 30–50 глосс, обработку неизвестных лексем и визуализацию результата. Нефункциональные требования определяют лимиты производительности (задержка ≤ 200 мс, работа ≥ 15 FPS на CPU уровня Intel i5/Ryzen 5, потребление ОЗУ ≤ 2 ГБ), локальность обработки данных, соответствие базовым рекомендациям WCAG 2.2 и фиксацию версий для воспроизводимости. Обоснован выбор технологического стека: MediaPipe Hands для детекции ключевых точек, BiLSTM для моделирования временных зависимостей, TensorFlow Lite для оптимизации инференса и Streamlit для быстрого прототипирования интерфейса.

Во второй главе описана архитектура приложения и конвейер обработки видеопотока. Система спроектирована по модульному принципу, что обеспечивает изоляцию этапов захвата, извлечения признаков, классификации и пользовательского взаимодействия. Входной видеопоток буферизуется в кольцевой очереди, после чего кадры передаются в детектор MediaPipe Hands, извлекающий 21 точку на каждую кисть с координатами (x, y, z) и меткой уверенности. Использование скелетного представления снижает размерность входных данных на несколько порядков и устраняет зависимость от фона и освещения.

Процесс предобработки включает двухэтапную нормализацию: пространственную (пересчёт координат относительно положения запястья для инвариантности к масштабу) и темпоральную (линейную интерполяцию всех последовательностей до фиксированной длины). Отсутствующие данные помечаются маркером маскирования, что позволяет нейронной сети игнорировать невалидные кадры. Предобработанные массивы сохраняются в бинарный формат .npy , что гарантирует детерминированность экспериментов и ускоряет последующие итерации обучения.

В качестве классификатора выбрана архитектура двунаправленной рекуррентной сети BiLSTM. Модель состоит из слоя маскирования, двух LSTM-слоёв (128 и 64 нейрона), регуляризационного слоя (Dropout = 0.3) и завершающего плотного слоя с функцией активации softmax. Обучение проводилось с использованием оптимизатора Adam, callbacks ReduceLROnPlateau и EarlyStopping. Ключевым методологическим решением стало строгое разделение данных по user_id : обучающая и валидационная выборки формировались исключительно из официального набора train датасета Slovo, тестовая выборка соответствовала набору test . Это исключает утечку данных и обеспечивает объективную проверку обобщающей способности модели на новых исполнителях.

После обучения модель конвертируется в формат TFLite с применением пост-тренировочной квантовки. Программная реализация модуля перевода включает загрузку оптимизированной модели, выделение тензоров ввода/-вывода и серию запусков для бенчмаркинга. Интерфейс прототипа реализован на Streamlit и предоставляет возможности загрузки видео, записи с веб-камеры, отображения распознанной глоссы с указанием уверенности и информативных сообщений об ошибках. В интерфейсе зафиксирован этический дисклеймер о прототипном статусе системы и локальной обработке данных.

В третьей главе представлена методика тестирования и результаты экспериментальной оценки разработанного прототипа. Экспериментальный контур спроектирован с учётом принципов научной воспроизводимости и строгой изоляции данных. Для проведения исследований использовано сбалансированное подмножество открытого датасета Slovo, содержащее 42 частотные

бытовые и учебные глоссы РЖЯ. Изначальный отбор включал 1220 видеозаписей, после применения фильтрации и ограничения до 150 примеров на класс в обучающей выборке итоговый объём составил 970 видео. Разделение данных выполнено по идентификаторам участников (user-based split), что исключает пересечение пользователей между множествами и предотвращает утечку данных (data leakage) из-за схожести антропометрии и стиля исполнения. Финальная разбивка выглядит следующим образом: обучающая выборка – 44 пользователя (839 видео), валидационная – 9 пользователей (85 видео), тестовая – 10 пользователей (46 видео). В качестве критериев качества классификации использованы Accuracy, F1-weighted, Precision, Recall и матрица ошибок. Выбор F1-weighted обусловлен мультиклассовым характером задачи и неравномерной представленностью глосс, что делает данную метрику более информативной по сравнению с простой точностью. Для оценки вычислительной эффективности фиксировались задержка инференса, пропускная способность (FPS), средняя загрузка CPU и объём потребляемой оперативной памяти.

Результаты классификации на отложенной тестовой выборке показали следующие агрегированные значения: Accuracy = 0.0270 (2.70%), F1-weighted = 0.0014. Данные показатели находятся на уровне случайного угадывания для 42 классов (≈2.38%), что свидетельствует об отсутствии устойчивого разделения признаков в текущей конфигурации. Анализ динамики обучения выявил существенный разрыв между кривыми обучающей и валидационной потерь: обе демонстрируют медленное снижение, однако абсолютные значения остаются высокими (train loss ≈3.72, val loss ≈3.73 на последних эпохах). Точность на обучающей выборке колебалась в пределах 1.5–2.5%, тогда как на валидационной приближалась к нулю, что указывает на недостаточную ёмкость выборки для эффективной настройки весов модели. Визуализация матрицы ошибок подтвердила склонность сети к предсказанию одного доминирующего класса, при котором остальные 41 глосса практически игнорируются. Основными причинами низких метрик признаны: ограниченный объём обучающих данных (в среднем ≈20 видео на класс после строгого разделения по пользователям), избыточная сложность архитектуры BiLSTM (≈436 тыс. параметров) относительно размера выборки, высокая кинематическая схо-

жесть отдельных глосс (особенно числительных, таких как «одиннадцать», «шестнадцать», «восемьдесят», отличающихся лишь темпом или размахом) и отсутствие учёта не-мануальных маркеров (мимики, положения корпуса, ориентации ладони). Несмотря на скромные классификационные результаты, эксперимент подтвердил полную техническую работоспособность конвейера: этапы извлечения признаков, пространственно-временной нормализации, инференса и локального выполнения на CPU осуществлялись корректно, детерминированно и воспроизводимо.

Бенчмаркинг, проведённый на оптимизированной модели в формате TensorFlow Lite, продемонстрировал соответствие ключевым нефункциональным требованиям локального развёртывания. Применение посттренировочной квантовки сократило объём файла весов на 88.7% (с 5.05 МБ до 0.57 МБ), что минимизирует требования к оперативной памяти и упрощает распространение решения. Средняя задержка инференса составила 107.4 мс, что укладывается в установленный лимит ≤ 200 мс и обеспечивает приемлемую отзывчивость в интерактивных сценариях. Пропускная способность достигла 9.3 FPS при целевом показателе 15 FPS; данный результат является достаточным для пакетной обработки загруженных видео, однако требует дополнительной оптимизации для стабильной работы в режиме реального времени. Средняя загрузка CPU зафиксирована на уровне 88.7%. Система функционирует полностью автономно: отсутствуют фоновые сетевые запросы, биометрические данные обрабатываются исключительно на стороне клиента, а интерфейс спроектирован с учётом базовых принципов доступности (WCAG 2.2), включая достаточную контрастность, клавиатурную навигацию и информативные сообщения об ошибках. В интерфейсе закреплён этический дисклеймер, исключающий применение прототипа в юридически или медицински значимых ситуациях.

Сформулированы системные ограничения и детализированы перспективы развития. Для преодоления выявленных узких мест предложены следующие направления оптимизации: расширение обучающей выборки до 50–100 примеров на класс с применением техник аугментации (временное масштабирование, добавление гауссова шума к координатам, случайное удаление кадров для имитации сбоев детектора); сокращение размерности LSTM-слоёв (до 64

и 32 нейронов) и усиление регуляризации (Dropout ≤ 0.5) для выравнивания соотношения параметров и данных; внедрение модулей MediaPipe Pose и Face Mesh для учёта пространственной организации знаков и мимической просодии; вычисление кинематических производных (скорость, ускорение углов суставов) как дополнительных временных признаков; переход к метрическим подходам (Siamese Networks, Prototypical Networks) и методам few-shot learning, адаптированным для работы с малыми выборками, а также реализация иерархической классификации. Для достижения целевой частоты 15 FPS рекомендуется снизить параметр TARGET_LENGTH до 15–20 кадров и применить INT8-квантование при конвертации модели. Сформулированные пути оптимизации позволяют рассматривать текущий прототип как воспроизводимый инженерный задел, закрывающий задачи локальной обработки, оптимизации инференса и защиты данных, что создаёт методическую основу для последующих исследований в области инклюзивных интерфейсов.

Заключение. В ходе выполнения выпускной квалификационной работы была достигнута поставленная цель: разработан и экспериментально оценён прототип локальной системы двунаправленного перевода между русским жестовым языком и русским языком. Система реализована с архитектурной асимметрией, где основное исследовательское внимание уделено направлению «жесты — текст» на базе последовательностного моделирования. Прототип оптимизирован для работы на центральных процессорах общего назначения без зависимости от облачных сервисов и прошёл строгую валидацию по идентификаторам участников.

Решение сформулированных задач позволило получить следующие результаты:

- Подготовлен детерминированный конвейер предобработки. На базе открытого датасета Slovo выделено подмножество из 42 частотных глосс. Разработан автоматизированный пайплайн, включающий извлечение скелетных признаков (MediaPipe Hands), пространственную нормализацию относительно запястья, темпоральное выравнивание последовательностей и сериализацию в структурированные бинарные файлы. Разделение выборки выполнено строго по user_id , что исключает утеч-

ку данных и обеспечивает корректную оценку обобщающей способности.

- Адаптирована и обучена последовательностная модель. В качестве классификатора применена архитектура BiLSTM, учитывающая двунаправленный временной контекст. Обучение сопровождалось регуляризацией и ранней остановкой. Низкие итоговые метрики (Accuracy 2.70%, F1-weighted 0.0014) обусловлены объективными факторами: строгим разделением по пользователям, создающим разрыв в антропометрии и стиле исполнения; малым объёмом обучающих примеров на класс; и высокой сложностью архитектуры относительно размера выборки. Несмотря на это, доказана техническая работоспособность пайплайна и способность модели обрабатывать скелетные данные в заданных условиях.
- Проведена оптимизация и экспорт модели. Обученная модель конвертирована в формат TensorFlow Lite с пост-тренировочной квантовкой, что сократило её объём на 88.7% (до 0.57 МБ). Бенчмаркинг на CPU подтвердил соблюдение нефункциональных требований: средняя задержка составила 107.4 мс, пропускная способность – 9.3 FPS, полная автономность обработки биометрических данных обеспечена.
- Разработан пользовательский интерфейс. На базе Streamlit создан веб-интерфейс, реализующий загрузку видео, отображение результатов с указанием уверенности модели и обработку граничных случаев. Интерфейс содержит этический дисклеймер и соответствует базовым принципам доступности.

Практическая значимость работы заключается в создании функционального, воспроизводимого прототипа, пригодного для использования в образовательных сценариях, бытовом взаимодействии и как аналитический инструмент для сбора размеченных данных. Разработанный конвейер (извлечение признаков — предобработка — обучение — экспорт в TFLite) может служить методической основой для расширения словарного запаса и адаптации системы под другие сценарии использования. Все экспериментальные артефакты зафиксированы, узкие места определены, что делает работу ценным заделом для последующих исследований в области инклюзивных технологий.

Перспективы развития системы включают:

- Расширение обучающей выборки и применение аугментации данных. Переход к более лёгким архитектурам или метрическим подходам (few-shot learning).
- Интеграцию модулей захвата позы и мимики.
- Вычисление кинематических производных координат.
- Оптимизацию инференса для достижения стабильных 15 FPS и переход к моделям непрерывного распознавания потока с использованием функции потерь CTC.

Результаты работы подтверждают техническую реализуемость легковесных систем распознавания жестового языка на стандартном оборудовании при соблюдении принципов локальности, воспроизводимости и защиты данных.