МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

# РАЗРАБОТКА НЕЙРОСЕТИ ОБРАБОТКИ АУДИОСИГНАЛА РАЗЛИЧНЫХ МУЗЫКАЛЬНЫХ ИНСТРУМЕНТОВ В ВЕБ-ПРИЛОЖЕНИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Мыльниковой Анастасии Сергеевны

Научный руководитель:
Старший преподаватель _____ Е.Е. Лапшева

подпись, дата

Зав. кафедрой:
Доцент, к.ф.-м.н. _____ М.В. Огнева

подпись, дата

Саратов 2026

# ВВЕДЕНИЕ

**Актуальность темы.** Реккурентные нейросети способны анализировать большое количество данных, упорядоченных по времени. Такие системы используют сложные алгоритмы машинного обучения, которые можно применить для синтеза звука, создания мелодий, гармоний и других элементов музыкальной композиции.

Обработка различных музыкальных инструментов с помощью нейросетей является актуальной задачей, так как искусственный интеллект влияет на все сферы деятельности и может быть применим на практике в современной музыкальной индустрии. В условиях стремления к автоматизации рутинных операций нейросетевые решения способны стать мощным инструментом для повышения эффективности аудиопроизводства. Существующие коммерческие решения, такие как Audo Studio и LALAL.AI, сосредоточены на решении узкоспециализированных задач. Прямых аналогов, представляющих собой приложение для адаптации звучания заранее записанных инструментов с использованием ML-сервиса (machine learning) выявлено не было. Данный факт подтверждает новизну предлагаемого подхода и определяет нишу для разрабатываемого проекта.

**Цель бакалаврской работы** – разработка сервиса, предоставляющего основному приложению нейросетевые модели для обработки аудиодорожек различных музыкальных инструментов и вокала. В отличие от существующих решений, ориентированных на генерацию музыки или разделение дорожек, предлагаемая система направлена на трансформацию звучания инструментов с сохранением их идентичности и артикуляции.

Поставленная цель определила **следующие задачи**:

1. Изучить возможные аналоги данного приложения среди уже существующих.
2. Исследовать существующие подходы к использованию рекуррентных нейронных сетей для обработки временных последовательностей.

3. Изучить методы спектрального анализа аудиосигналов и влияние оконных функций на качество представления данных.

4. Изучить представление музыки для компьютера.

5. Выбрать подходящую архитектуру модели нейросети для практической части.

6. Рассмотреть виды существующих эффектов, применяемых в аранжировке и сведении музыки.

7. Изучить библиотеки в Python подходящие для работы с нейросетями и аудио-сигналами.

8. Подготовить консистентный датасет для каждого инструмента.

9. Обучить на датасетах модель обрабатывать инструменты.

10. Создать сервис на фреймворке FastAPI с брокером сообщений и объектным хранилищем, чтобы интегрироваться с основным приложением.

**Методологические основы** разработки нейросети обработки аудио сигнала различных музыкальных инструментов в веб-приложении представлены в работах Ф. Харриса, Х. Пурвинса и соавторов, Л.-Д. Жимон, В. Чжан и соавторов, Е. Рулько, У. Аурелио, А. Хашим.

**Теоретическая значимость бакалаврской работы.** заключается в обосновании выбора модифицированной архитектуры UNet (с механизмами внимания и остаточными блоками) для задачи преобразования тембра музыкальных инструментов с сохранением структуры спектра, а также в систематизации критериев выбора оконных функций (на примере окна Хэннинга) для спектрального анализа в задачах машинного обучения.

**Практическая значимость бакалаврской работы.** состоит в разработке и внедрении готового к использованию MLсервиса, интегрированного с вебприложением через асинхронную очередь сообщений Apache Kafka и объектное хранилище MinIO. В процессе работы подготовлены консистентные датасеты для пяти типов инструментов, обучены специализированные модели, разработаны скрипты их применения. Предложенные подходы могут быть использованы в профессиональной

звукорежиссуре для автоматизации обработки и стилизации звука, а также в образовательных целях.

**Структура и объём работы**. Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и четырех приложений. Общий объем работы – 99 страниц, из них 42 страницы – основное содержание, включая 10 рисунков, список использованных источников информации – 24 наименований.

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Первый раздел «Теоретические основы обработки аудио сигнала и применения нейросетей для обработки»** посвящён комплексному анализу предметной области и теоретическому обоснованию проектных решений.

В начале раздела рассматриваются существующие коммерческие и исследовательские решения, которые можно классифицировать по трём направлениям. Первое направление – автоматический мастеринг и улучшение звука: облачные сервисы LANDR и iZotope Ozone 11, использующие нейросетевые модели для анализа трека и подбора оптимальных параметров обработки (эквалайзер, компрессия, лимитирование), однако они не выполняют аранжировку отдельных инструментальных дорожек. Второе направление – генеративные системы, создающие музыку «с нуля»: AIVA (основана на глубоких рекуррентных архитектурах и обучена на партитурах классической музыки), OpenAI Jukebox (сочетает VQVAE и авторегрессионный трансформер для генерации музыки с вокалом), Suno AI (кросс-модальная генерация по текстовому описанию) и Riffusion (адаптация диффузионной модели Stable Diffusion для работы со спектрограммами). Третье направление – инструменты разделения и очистки аудиосигналов: LALAL.AI (модель Rocknet для выделения вокала и инструментов) и Audo Studio (удаление шума и улучшение речи). Анализ показывает, что прямых аналогов, ориентированных именно на трансформацию тембра заранее записанных инструментов с сохранением их

идентичности и артикуляции, не существует, что обосновывает новизну разрабатываемого подхода и определяет нишу для проекта.

Далее выполнен детальный обзор архитектур нейронных сетей, применяемых в аудиообработке. Рекуррентные нейронные сети (RNN) и их модификации LSTM и GRU исторически были одними из первых глубоких архитектур для работы с последовательными данными благодаря способности накапливать контекст из предыдущих шагов. Однако базовые RNN страдают от проблемы затухающего градиента; LSTM решает эту проблему за счёт ячейки памяти и системы ворот (forget gate, input gate, output gate), а GRU является упрощённой версией с двумя вентилями и меньшим числом параметров. В рамках данной работы были проведены первоначальные эксперименты с GRU-архитектурой на мел-спектрограммах, которые показали неудовлетворительные результаты: восстановленный сигнал характеризовался высоким уровнем шума, размытием спектральных компонент и потерей тональной чистоты. Анализ причин выявил фундаментальные ограничения рекуррентных архитектур для данной задачи: они обрабатывают спектрограмму как одномерную последовательность временных кадров, теряя пространственные (частотные) корреляции; не допускают параллельной обработки, что критично при высокоразмерных STFT-спектрограммах; не имеют встроенного механизма пространственной локализации для фокусировки на конкретных частотных областях. Кроме того, мел-спектрограмма как входное представление оказалась непригодной из-за необратимой потери высокочастотной информации вследствие нелинейного сжатия частотной оси. Трансформеры, основанные на selfattention, демонстрируют впечатляющие результаты в задачах моделирования глобальных зависимостей, однако их квадратичная сложность относительно длины последовательности делает их крайне ресурсоёмкими для длинных аудиозаписей, а для задач локальной структуры спектра глобальное внимание может быть избыточным. Свёрточные нейронные сети эффективно извлекают локальные паттерны из двумерных спектральных

представлений, но стандартные классификационные CNN не решают задачи преобразования сигнала той же размерности необходима архитектура типа «кодировщик-декодировщик». Генеративно-состязательные сети и диффузионные модели позволяют синтезировать высококачественный звук, но направлены прежде всего на задачи генерации, а не трансформации существующего сигнала, и отличаются нестабильностью обучения либо высокой вычислительной сложностью.

На основе проведённого анализа и результатов экспериментов был сделан вывод о необходимости одновременной смены как архитектуры, так и входного представления. В качестве нового представления сигнала выбран STFT (кратковременное преобразование Фурье), сохраняющий полную линейную частотную информацию и допускающий точное восстановление сигнала через обратное преобразование. В качестве архитектуры выбрана UNet свёрточная модель типа «кодировщик-декодировщик» с симметричными skipсоединениями, рассматривающая STFTспектрограмму как двумерное изображение. Исследования подтверждают, что модели, работающие с двумерным представлением спектрограммы, стабильно превосходят чисто рекуррентные подходы в задачах разделения и восстановления источников. В работе реализована модифицированная UNet, включающая: кодировщик из трёх уровней с последовательным уменьшением разрешения вдвое и увеличением числа каналов, bottleneck из двух Residualблоков, позволяющих обучать более глубокие сети без деградации точности за счёт прямого прохождения градиента через skipсвязи; декодировщик с транспонированными свёртками, восстанавливающий исходное разрешение, Attention Gates в skipсоединениях вместо простой конкатенации, которые динамически вычисляют карту важности с помощью сигмоидной функции, позволяя модели фокусироваться на акустически активных зонах спектрограммы и подавлять шумовые области, финальный свёрточный слой с активацией Sigmoid для формирования одноканальной спектрограммы-маски. Архитектура оптимизирована для работы на

устройствах с Apple Silicon за счёт использования LeakyReLU (коэффициент 0.2) вместо более тяжёлых активаций и ограничения глубины тремя уровнями, что снижает потребление памяти без существенной потери качества.

Значительное внимание в разделе уделено методам спектрального анализа. Рассмотрено кратковременное преобразование Фурье (STFT), определяемое как свёртка сигнала с оконной функцией, сдвигаемой по времени. Подробно проанализировано влияние различных оконных функций на качество спектрального представления: прямоугольное окно (самая узкая ширина главного лепестка $4\pi/N$, наилучшее частотное разрешение, но наибольший уровень боковых лепестков около 13 дБ, что приводит к сильной спектральной утечке); треугольное окно (уровень боковых лепестков около 25 дБ, промежуточное положение); окна с поднятым косинусом – Хэннинга ($0.5\cdot(1-\cos(2\pi n/(N-1)))$, уровень боковых лепестков около 31 дБ, ширина главного лепестка $8\pi/N$) и Хэмминга ($0.54-0.46\cdot\cos(2\pi n/(N-1))$, уровень первого бокового лепестка около 43 дБ, но медленный спад последующих); окна косинусной суммы – Блэкмана (уровень боковых лепестков около 58 дБ, ширина главного лепестка $12\pi/N$) и Блэкмана-Наттала (уровень боковых лепестков около 98 дБ, ширина главного лепестка $14\pi/N$). Для выбора оптимальной оконной функции в проекте применена методика, основанная на двух критериях: требуемый динамический диапазон анализа (уровень боковых лепестков должен быть ниже динамического диапазона сигнала) и необходимое частотное разрешение (ширина главного лепестка определяет минимальное расстояние между разрешаемыми частотами). Аудиодорожки музыкальных инструментов в проекте имеют динамический диапазон 70 дБ (от +10 дБ до −60 дБ), что теоретически требует окна с уровнем боковых лепестков ниже −70 дБ, такого как Блэкмана-Наттала. Однако его ширина главного лепестка ($K\approx12$) при частоте дискретизации 48 кГц и требуемом разрешении $\Delta f=10$ Гц потребовала бы размера окна $N \geq (12\cdot48000)/10 = 57600$ отсчётов (1.2 секунды), что полностью нивелирует временное разрешение и

неприемлемо для анализа динамических музыкальных событий. В результате выбран компромиссный вариант – окно Хэннинга с размером N=2048 (около 43 мс) и перекрытием 75% (hop=512), обеспечивающее частотное разрешение $\Delta f \approx (8 \cdot 48000)/(2\pi \cdot 2048) \approx 29.8$ Гц, достаточное для анализа большинства музыкальных гармоник, при удовлетворительном подавлении спектральной утечки для практических задач. Этот выбор также соответствует общепринятой практике в области аудиоанализа и широко используется в библиотеках для вычисления спектральных представлений.

В завершении раздела сформулированы итоговые требования к архитектуре модели, параметрам STFT и системе предобработки, что создаёт теоретическую основу для практической реализации, описанной во втором разделе.

**Второй раздел «Практическая реализация проекта»** посвящен реализации описывает полный цикл разработки программного решения. Подготовка датасета заняла около года и включала сбор более 150 Гб живых записей пяти типов инструментов (акустическая гитара, электрогитара, бас-гитара, клавишные, вокал) с помощью программы Reaper, а также их последующую сегментацию на 15-секундные фрагменты, нормализацию громкости и обеспечение консистентности данных (одинаковая частота дискретизации 48 кГц, единообразная обработка). Для каждого инструмента подготовлены парные наборы «сырых» и «обработанных» записей. В качестве архитектуры выбрана улучшенная UNet с тремя уровнями кодирования/декодирования, Attention Gates в skip-соединениях и Residualблоками в bottleneck, оптимизированная для работы на устройствах с Apple Silicon (использование LeakyReLU). Для обучения применена комбинированная функция потерь, включающая L1-компоненту для точности амплитуды, SSIM-компоненту для сохранения структурного сходства спектрограмм и edgекомпоненту для резкости спектральных границ. Оптимизатор Adam с начальной скоростью обучения 1e3 и scheduler StepLR использовались на протяжении до 50 эпох с ранней остановкой. Для каждого

инструмента обучена отдельная модель, а постобработка настроена с учётом акустических особенностей: для акустической гитары применены спектральное сглаживание и гейтинг с частотной коррекцией; для басгитары фильтр низких частот 5 кГц, для электрогитары агрессивное шумоподавление и усиление низких частот, для клавишных минимальная обработка с увеличенным перекрытием сегментов для сохранения длительного затухания. В заключительной части раздела описана разработка MLсервиса на фреймворке FastAPI, интегрированного с основным приложением через брокер сообщений Apache Kafka и объектное хранилище MinIO. Сервис использует паттерн «Стратегия» для выбора модели в зависимости от типа инструмента и реализует overlapadd метод для обработки длинных файлов. Система обеспечивает уведомления о прогрессе через SignalR. Представлены результаты в виде спектрограмм, демонстрирующих качественное восстановление спектральной структуры для большинства инструментов. Выводы по разделу подтверждают эффективность выбранного подхода и применимость системы в профессиональной звукорежиссуре, а также намечены пути дальнейшего развития: улучшение модели для электрогитары, расширение набора инструментов и добавление дополнительных эффектов.

# ЗАКЛЮЧЕНИЕ

Выполнение данной работы позволило реализовать комплексное решение для обработки аудиосигналов музыкальных инструментов на основе методов глубокого обучения. Была разработана и внедрена полная цепочка обработки: от сбора и подготовки данных до развёртывания модели в виде микросервиса, интегрированного с основным приложением.

В ходе исследования были изучены особенности спектрального анализа аудиосигналов, обосновано применение кратковременного преобразования Фурье (STFT) в сочетании с оконной функцией Хэннинга, что позволило снизить спектральные искажения и повысить устойчивость модели к артефактам на границах фрагментов.

В качестве основы архитектуры была выбрана модифицированная U-Net, адаптированная для работы с частотно-временными представлениями аудиосигналов. Эта архитектура, включающая механизмы внимания, остаточные блоки и skip-соединения, показала высокую эффективность в восстановлении структуры спектрограммы, что критично для сохранения тембровых характеристик инструментов.

Для каждой группы инструментов были подготовлены специализированные датасеты и обучены отдельные модели, что позволило учесть их акустические особенности. Использование комбинированной функции потерь, включающей L1, SSIM и edge-компоненты, способствовало достижению баланса между точностью восстановления и сохранением деталей спектра.

Был реализован сервис на фреймворке FastAPI, обеспечивающий приём аудиофайлов, их спектральное преобразование, обработку моделью и возврат результата. Сервис интегрирован с основным приложением через брокер сообщений Apache Kafka и объектное хранилище MinIO, что обеспечивает отказоустойчивость, масштабируемость и асинхронную обработку задач.

Таким образом, в ходе работы была создана эффективная система обработки аудиосигналов, применимая в профессиональной звукорежиссуре.

Полученные результаты подтверждают перспективность использования U-Net-архитектур в задачах аудиопроцессинга и создают основу для дальнейшего улучшения модели, расширения набора поддерживаемых инструментов и интеграции дополнительных эффектов.

**Основные источники информации:**


1. Harris F.J. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform // Proceedings of the IEEE. – 1978. – Vol.66, No. 1.
2. Purwins H., Li B., Virtanen T., Schlüter J., Chang S., Sainath T. Deep Learning for Audio Signal Processing // Journal of Selected Topics of Signal Processing. – 2019. – Vol. 13, № 2. – P. 206–219.
3. Jimon L.-D. Deep Learning Approaches in Audio Processing: Recent Advances and Challenges // Технический университет КлужНапоки. – 2025. – Vol. 65, № 1.
4. Zhang W., He C., Cao Y., Xu S., Wang M. Two-Stage Unet with Gated-Conv Fusion for Binaural Audio Synthesis // Sensors. – 2025. – Т. 25, № 6. – С. 1790.
5. Window Functions and Their Applications in Signal Processing – researchgate.net.
6. Rulko E. Applying attention U-Net with PyTorch architectural add-ons for extensive hyperparameter search with Weights & Biases for area of visibility prediction based on terrain – researchgate.net.
7. The Fundamentals of Signal Analysis / HP Memory Project.
8. Aurelio U. Digital Audio Processing: Fundamentals and Applications. – Cham: Springer, 2023.
9. Hashim A. Machine Learning Engineering with Python: Production-Grade Pipeline Automation and MLOps for AI. – Birmingham: Packt Publishing, 2022.