МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра математической кибернетики и компьютерных наук

РАЗРАБОТКА АЛГОРИТМА ГЕНЕРАЦИИ ИГРОВЫХ КАРТ ДЛЯ РИТМ-ИГРЫ С ИСПОЛЬЗОВАНИЕМ ТРАНСФОРМЕРНЫХ МОДЕЛЕЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование информационных систем
факультета КНиИТ
Мазанова Максима Александровича

Научный руководитель
к. ф.-м. н., доцент _____ С. В. Папшев

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Современная разработка видеоигр сталкивается с ростом стоимости производства контента. В ритм-играх основа геймплея — это чарт, то есть последовательность объектов, строго синхронизированная с музыкой. Ручное создание качественного чарта требует высокой ритмической точности и понимания эргономики геймдизайна, занимая десятки часов. Существующие методы процедурной генерации не всегда обеспечивают точную синхронизацию и гибкое управление сложностью. Разработка автоматизированного, управляемого метода генерации чартов позволяет снизить трудозатраты и ускорить прототипирование.

Целью магистерской работы является разработка метода автоматической генерации игровых карт для ритм-игры osu!mania с использованием мультимодальной трансформерной модели, способной учитывать параметры сложности и акустические признаки композиции.

Поставленная цель определила **задачи**:

1. Проанализировать существующие подходы процедурной генерации символьной музыки и ритмических последовательностей и оценить их применимость к задаче автоматического маппинга в ритм-игре *osu!mania*.
2. Собрать набор данных на основе пользовательских чартов и разработать конвейер предварительной обработки для извлечения спектральных характеристик аудиосигнала и нормализации игровой разметки.
3. Разработать схему токенизации для перевода пространственно-временной структуры чарта в дискретную последовательность, пригодную для обучения авторегрессионной модели.
4. Спроектировать алгоритмы извлечения векторов признаков, описывающих макро- и микрохарактеристики сложности композиции, для управления параметрами генерации.
5. Разработать архитектуру мультимодальной трансформерной модели с механизмом перекрестного внимания, обеспечивающую интеграцию акустических признаков и условных параметров в процессе генерации.
6. Обучить генеративную модель, подобрать оптимальную стратегию оптимизации и реализовать мониторинг ключевых метрик качества.
7. Выполнить количественную и качественную оценку сгенерированных чартов с использованием F1-метрик и визуального анализа.

Апробация работы. Основные результаты исследования были представлены на конференциях: *Presenting Academic Achievements to the World XVI* (2025) и *Студенческая научная конференция СГУ* (2026). По теме работы опубликованы две статьи в сборниках материалов конференций.

Методологические основы работы с ключевыми фразами представлены в работах C. Donahue, Z. C. Lipton и J. McAuley; Y. Liang, W. Li и K. Ikeda; E. Halina и M. Guzdial; J. J. Yi, S. Lee и K. Lee; A. Takada и др.

Теоретическая значимость магистерской работы. Проведена формализация создания ритмических уровней как задачи обусловленной авторегрессионной генерации. Разработан подход к извлечению дискретных признаков для описания макроструктуры чарта. Предложена двухуровневая система управления стилистикой в трансформерах для ритм-игр с вертикальной прокруткой с помощью линейной модуляции признаков и локальных префиксных токенов.

Практическая значимость магистерской работы. Разработан масштабируемый конвейер для автоматического синтеза файлов .osu из аудиосигналов. Сформирован и очищен масштабный набор данных, включающий более 19 тысяч пользовательских чартов, готовый к последующим исследованиям. Разработанная система может применяться как интеллектуальный инструмент маппера, способный за несколько минут генерировать качественный «черновик» заданного уровня сложности, минимизируя время рутинной разметки.

Научная новизна работы.

1. Впервые для ритм-игр с вертикальной прокруткой предложена двухуровневая архитектура условной генерации, интегрирующая механизмы глобальной линейной модуляции признаков и обучаемых префиксных эмбеддингов, что позволяет контролировать макроструктуру и локальную динамику паттернов.
2. Разработан метод динамической обусловленности для инференса: эвристическая аппроксимация акустических признаков с экспоненциальным сглаживанием обеспечивает генерацию согласованных чартов без эталонной разметки.
3. Схема токенизации адаптирована под специфику непрерывных объектов (удержаний), а введенное разделение метрик на колонко-зависимые (CW-F1) и независимые (CA-F1) формализует границу между точностью алгоритма и субъективностью авторского стиля.

Положения, выносимые на защиту.

1. Архитектура авторегрессионного трансформера с перекрестным вниманием к мел-спектрограммам эффективно извлекает ритмическую сетку, превосходя качество предсказания без опоры на аудиоконтекст.
2. Двухуровневая интеграция управляющих векторов обеспечивает детерминированный контроль над интенсивностью и семантикой чарта (доли паттернов, удержаний) на глобальном и локальном уровнях композиции.
3. Разработанный метод инференса на основе акустических эвристик и экспоненциального сглаживания позволяет автономно генерировать макроструктуру уровня с высокой корреляцией плотности (до 0.75 по Пирсону) относительно работ экспертов.
4. Предложенное разделение метрик позиционирования доказывает, что модель достигает колонко-независимой точности (CA-F1 = 85.2%) на уровне мировых SOTA-решений, тогда как колонко-зависимая точность (CW-F1) ограничена субъективностью геймдизайна.

Структура и объем работы. Магистерская работа состоит из перечня определений, обозначений и сокращений, введения, трех разделов, заключения, списка использованных источников и трех приложений. Общий объем работы — 106 страниц, из них 74 страницы — основное содержание, включая введение, основные разделы и заключение. Работа содержит 25 рисунков, 6 таблиц, цифровой носитель в качестве приложения, список использованных источников из 37 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретическая основа и проблематика» посвящен анализу предметной области, обзору существующих методов генерации контента, формализации задачи и рассмотрению теоретических основ применяемых алгоритмов.

Игровой процесс ритм-игры osu!mania. Рассмотрены механики взаимодействия с контентом (одиночные нажатия, удержания). Описан процесс создания чартов, влияние плотности (NPS) и темпа (BPM) на интенсивность геймплея. Подробно классифицированы структурные паттерны режима 4K: одиночные последовательности, аккорды, джеки, трели, а также сложные паттерны удержаний. Сформулированы технические (ритмическая точность, структурная непротиворечивость) и эргономические (баланс нагрузки, ограничение плотности) критерии качества маппинга.

Обзор существующих методов генерации в ритм-играх. Проведен анализ эволюции подходов: от покадровой бинарной классификации (модели DDC, C-BLSTM) до современных архитектур совместного моделирования (Taiko-Nation, GenéLive!). Рассмотрены SOTA-решения на базе трансформеров с выравниванием спектрограмм по границам тактов. Выявлены фундаментальные ограничения существующих систем: упрощенная скалярная оценка сложности, отсутствие локального контекста для учета смены музыкальной динамики и неспособность отражать многомерную стилистику конкретного автора.

Теоретические и математические основы применяемых методов. Аргументирован выбор логарифмических мел-спектрограмм в качестве входного представления аудио. Изложены математические основы сверточных нейронных сетей (CNN) для извлечения локальных частотно-временных паттернов. Подробно описана архитектура трансформера: механизм масштабированного скалярного внимания, каузальное маскирование в авторегрессионном декодере и конфигурация Pre-Norm для глубоких сетей. Рассмотрены математические методы обусловленной генерации: покомпонентная аффинная линейная модуляция (FiLM), масштабирующая и сдвигающая активации, а также метод обучаемых префиксных эмбеддингов для внедрения локального контекста.

Второй раздел «Реализация модели генерации» посвящен проектированию программного конвейера, обработке данных и детальному описанию архитектуры нейросетевой модели и процессу обучения.

Архитектура системы и модуль анализа. Разработан программный комплекс на языке Python (библиотеки $\mathrm{PyTorch}$, $\mathrm{librosa}$). Конвейер разделен на подготовку данных, обучение и инференс. Создан $\mathrm{BeatmapAnalyzer}$ — модуль, переводящий абстрактные концепции геймдизайна в строгие метрики. Для оценки локальной интенсивности используется скользящее окно шириной в одну секунду, обработка пересечений удержаний реализована алгоритмом заметающей прямой, а для оценки баланса рук применяется энтропия Шеннона.

Формирование и анализ датасета. Собрано 6145 наборов карт со статусами Ranked и Loved. После очистки от физических коллизий финальный объем составил 19617 чартов *osu!mania 4K*. Статистический анализ подтвердил высокое качество данных: 94% объектов привязаны к стандартным квантам ($1/1$, $1/2$, $1/4$), среднее отклонение от сетки составило всего 0.71 мс. Выявлена широкая дисперсия плотности (NPS от 0.77 до 34.81), средняя доля многонотных аккордов составила 36.7%, удержаний — 18.3%. Корреляционный анализ метрик доказал независимость структурных параметров, что обосновало их совместное использование в векторах признаков.

Токенизация и формирование векторов признаков. Разработана событийно-ориентированная токенизация. Введено относительное ритмическое разрешение 48 тиков на долю, что позволяет точно описывать как стандартные ($1/2$, $1/4$, $1/8$), так и триольные ($1/3$, $1/6$) доли без накопления ошибок округления. Алфавит включает 251 токен: маркеры сегмента, события нажатий NOTE_K и HOLD_K, сдвиги времени до 48 тиков TIME_SHIFT_N, длительности удержаний до 192 тиков DUR_N. Сформирована двухуровневая система векторов: глобальный тензор (6 макрохарактеристик) и локальный тензор сегмента (12 характеристик, включая физическую длительность в мс и флаги усечения $\mathrm{is\_tail}$).

Конвейер предобработки данных и архитектура модели. Аудио извлекается с частотой 22050 Гц, окном 2048 и шагом 512, формируя матрицы [80, 430] для четырехтактовых сегментов. Применяется Z-масштабирование признаков с логарифмическим сжатием $\mathrm{log1p}$ для асимметричных распределений (джеки, трели). Спроектирован условный трансформер (скрытая размерность 512, 8 голов, FFN 2048). Аудиоэнкодер состоит из 3 сверточных блоков (свертки 3x3, GELU, Max Pooling) и 3 слоев TransformerEncoder. Декодер (6 слоев) включает механизм FiLM на каждом слое для глобальной стилистики и конкатенацию

токена SEG_EMBED для локальной адаптации. Применяется связывание весов (*weight tying*) выходного слоя со словарем эмбеддингов.

Загрузчик данных, обучение и мониторинг. Для эффективной пакетной обработки реализован загрузчик данных с маскированием внимания и разделением выборки по уникальным идентификаторам наборов карт (исключена утечка данных между обучающей и валидационной выборками). Процесс обучения включает оптимизатор AdamW, планировщик скорости с линейным разогревом и косинусным затуханием, а также механизм накопления градиентов для увеличения эффективного размера батча. В процессе обучения вычисляются категориальная перекрестная энтропия, точность предсказаний (в том числе Top-3 Accuracy) и нормализованная энтропия распределения. Для оценки вклада аудиосигнала проводится аблационное тестирование. Реализован механизм ранней остановки и сохранения контрольных точек.

Конвейер генерации и анализ аудиосигнала. Для применения обученной модели к новым композициям разработан модуль AudioAnalyzer, извлекающий из аудиосигнала частоту атак, спектральный поток, автокорреляцию и RMS-энергию. Эти признаки масштабируются под выбранный пользователем профиль сложности (от Easy до Expert) и преобразуются в локальные векторы признаков с применением экспоненциального сглаживания для обеспечения плавных переходов между сегментами. Авторегрессионная генерация выполняется с использованием температурного масштабирования и стратегий усечения распределения (Top-k и Top-p), что позволяет балансировать между детерминированностью и разнообразием. Завершающие этапы конвейера включают декодирование последовательности токенов в игровые объекты, фильтрацию (устранение коллизий и микро-пауз) и экспорт в файл формата .osu.

Третий раздел «Экспериментальные исследования и оценка результатов» содержит описание условий эксперимента, анализ сходимости модели, количественную и качественную оценку сгенерированных чартов.

Настройка эксперимента. Обучение проводилось на видеокарте NVIDIA GeForce RTX 3090 (24 ГБ) с использованием PyTorch. Датасет разделен на обучающую (90%) и валидационную (10%) выборки с исключением утечки данных на уровне наборов карт. Основные гиперпараметры: скрытая размерность 512, 8 голов внимания, 3 слоя энкодера, 6 слоев декодера, скорость обучения 10^{-4}, эффективный размер батча 360, ранняя остановка с patience 20 эпох.

Обучение и сходимость. Процесс остановлен на 82-й эпохе, лучшая модель зафиксирована на 62-й эпохе. Значение функции потерь на валидационной выборке составило 1.469, перплексия — 4.306, Top-3 Accuracy — 94.98%. Анализ точности по типам токенов показывает: TIME_SHIFT — 85.0%, DUR — 69.4%, NOTE — 66.1%, HOLD — 55.4%. При маскировании аудиосигнала значение функции потерь ($\mathrm{zero\_loss}$) возрастало до 3.52, тогда как при использовании реальных акустических данных ($\mathrm{real\_loss}$) оно составляло 0.77. Зафиксированная разница на уровне 2.75 демонстрирует, что механизм перекрёстного внимания функционирует корректно: для выбора временных позиций игровых объектов и их типов нейронная сеть опирается на спектральные признаки музыкального сопровождения.

Оценка управляемости и количественный анализ. Для оценки позиционирования применена F1-мера с окном допуска ± 45 мс. Вычислялись колонконезависимая (CA-F1) и колонко-зависимая (CW-F1) метрики. Эксперимент по генерации композиции «POLKAMANIA» показал строгую аппроксимацию заданных целевых NPS и долей паттернов для всех профилей от Easy до Expert. Тестирование полного конвейера (8 композиций разных жанров) показало средний CA-F1 = 76.2% и CW-F1 = 36.3%. Стабильно высокая метрика CA-F1 подтверждает оптимальную ритмическую точность, в то время как низкий CW-F1 иллюстрирует субъективность ручного маппинга. Коэффициент корреляции Пирсона между профилями плотности генератора и человека достигает 0.75 на высоких сложностях.

Количественное сравнение с эталонной разметкой и SOTA-моделями. На валидационной выборке (66 653 сегмента) достигнут CA-F1 = 85.2%, CW-F1 = 43.1%. Полученные результаты сопоставимы с современными SOTA-решениями. Для задачи извлечения временных позиций метрика CA-F1 (85.2%) превосходит результаты архитектуры *GenéLive!* (80.2% при окне ± 50 мс) и находится на одном уровне с передовыми моделями *Yi et al.* (84.6% при ± 30 мс) и *Liang et al.* (84.3%). Колонко-зависимая точность (43.1%) превосходит показатели базовых нейросетевых моделей, таких как *AutoOsu* (13.6%), и согласуется с результатами *Liang et al.* (43.6%). Это подтверждает, что интеграция перекрестного внимания и условной генерации позволяет извлекать ритмическую сетку на уровне ведущих решений.

Качественный визуальный анализ. Сравнение сгенерированных и ори-

гинальных чартов показало, что модель корректно синхронизирует объекты с ритмическими акцентами, однако иногда завышает плотность, интерпретируя микро-колебания спектра как самостоятельные события (переход на шаг 1/4 вместо 1/2). Обнаружена алгоритмическая прямолинейность: при повторяющихся звуках модель генерирует джеки в одной колонке, тогда как человек чередует колонки для эргономичности. Наиболее существенный артефакт — генерация объектов в акустических паузах (авторегрессионные галлюцинации), вызванная инерцией целевых векторов плотности.

Обсуждение результатов. Разработанная модель подтверждает способность генерировать ритмически точные и физически выполнимые чарты, управляемые двухуровневой системой признаков. Выявленные ограничения (чувствительность к качеству профилирования, склонность к статистически безопасным паттернам, стохастичность) определяют позиционирование системы как инструмента ускоренного прототипирования, требующего последующей ручной доработки. Перспективные направления включают создание интеллектуального редактора для ручного задания признаков и адаптацию модели для других задач audio-to-sequence.

ЗАКЛЮЧЕНИЕ

В ходе выполнения магистерской работы изучены существующие подходы к процедурной генерации контента в ритм-играх, выявлены их ограничения (скалярное представление сложности, отсутствие локального контекста и механизмов управления стилистикой). На основе проведенного анализа сформулирована задача автоматического маппинга как обусловленной авторегрессионной генерации последовательностей.

В процессе исследования разработан и реализован полный конвейер обработки данных, включающий парсинг пользовательских чартов, извлечение логарифмических мел-спектрограмм, событийно-ориентированную токенизацию с разрешением 48 тиков на долю и формирование двухуровневых векторов признаков (глобального и локального) для управления генерацией.

Спроектирована и обучена мультимодальная трансформерная модель архитектуры «энкодер-декодер» с механизмами перекрестного внимания к аудиосигналу, линейной модуляции признаков (FiLM) для глобальной обусловленности и префиксных эмбеддингов для локальной адаптации. Для применения модели к новым композициям создан эвристический модуль анализа аудиосигнала, преобразующий акустические характеристики в управляющие векторы.

Результаты экспериментальных исследований подтверждают эффективность предложенного подхода. На валидационной выборке достигнут показатель колонко-независимой точности позиционирования CA-F1 = 85.2%. Сравнение с современными SOTA-решениями (*GenéLive!*, *AutoOsu*, модели *Liang et al.* и *Yi et al.*) подтвердило конкурентоспособность разработанной архитектуры: точность позиционирования объектов находится на уровне ведущих аналогов, а механизмы обусловленной генерации позволили превзойти базовые решения в задаче колонко-зависимого предсказания. Аблационное тестирование показало, что при маскировании аудиосигнала ошибка предсказания возрастает на 2.75 по сравнению с использованием реальных акустических данных, что доказывает эффективность механизма перекрёстного внимания. Оценка полного конвейера на восьми тестовых композициях продемонстрировала средний CA-F1 = 76.2%, а коэффициент корреляции Пирсона профилей плотности на высоких сложностях достигает 0.75, что свидетельствует о структурной согласованности с оригинальными чартами.

Таким образом, цель работы — разработка метода автоматической генера-

ции игровых карт для osu!mania с учетом параметров сложности и акустических признаков — достигнута. Все поставленные задачи выполнены. Разработанная система может применяться как инструмент ускоренного прототипирования уровней, сокращающий рутинные затраты на разметку.

Основные источники информации:

1. *Donahue C.*, *Lipton Z. C.*, *McAuley J.* Dance Dance Convolution // Proceedings of the 34th International Conference on Machine Learning (ICML 2017). Vol. 70. — Cambridge, MA, USA : PMLR, 2017. — P. 1039–1048.
2. *Liang Y.*, *Li W.*, *Ikeda K.* Procedural Content Generation of Rhythm Games Using Deep Learning Methods // Proceedings of the First IFIP TC14 Joint International Conference on Entertainment Computing and Serious Games (ICEC-JCSG). — Cham, Switzerland : Springer, 2019. — P. 134–145.
3. *Halina E.*, *Guzdial M.* TaikoNation: Patterning-focused Chart Generation for Rhythm Action Games // Proceedings of the 16th International Conference on the Foundations of Digital Games. — New York, NY, USA : ACM, 2021. — P. 1–10.
4. *Yi J. J.*, *Lee S.*, *Lee K.* Beat-Aligned Spectrogram-to-Sequence Generation of Rhythm-Game Charts // Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR 2023). — Canada : International Society for Music Information Retrieval, 2023.
5. GenéLive! Generating Rhythm Actions in Love Live! / A. Takada [et al.] // Proceedings of the 37th AAAI Conference on Artificial Intelligence. Vol. 37. — Palo Alto, CA, USA : AAAI Press, 2023. — P. 5266–5275.