УДК 004.056.57

МЕТОДИКА ОЦЕНКИ СТЕПЕНИ СХОЖЕСТИ АУДИОФАЙЛОВ С ПРИМЕНЕНИЕМ AMT-МОДЕЛИ

Ладыгин Павел Сергеевич, Неверов Андрей Игоревич

Алтайский государственный университет, Барнаул
pavel-ladygin@yandex.ru. andrey.neverov2003@mail.ru

A METHOD FOR ASSESSING AUDIO FILE SIMILARITY USING AN AMT MODEL

Ladygin Pavel S., Neverov Andrey I.

Altai State University, Barnaul
pavel-ladygin@yandex.ru. andrey.neverov2003@mail.ru

Аннотация. В статье рассматривается разработка и экспериментальная оценка нового метода создания цифровых отпечатков аудиофайлов на основе автоматической музыкальной транскрипции (AMT) для повышения эффективности выявления пиратского контента. В качестве методологической основы предлагается алгоритм, преобразующий аудиосигнал в символьное представление (MIDI) с помощью специально обученной нейросетевой модели «Google Onsets and Frames». Данная модель, архитектура которой включает два U-Net блока и слои BiLSTM, была обучена на датасете фортепианных записей MAPS. Для оценки схожести исходного и предсказанного MIDI-представления использовалось расстояние Левенштейна. Экспериментальные результаты выявили сильную зависимость точности модели от характера аудиоданных. Наивысшая точность (89,36%) была достигнута на полифонических фортепианных композициях, соответствующих обучающим данным, тогда как на монофонических и гитарных треках результаты были значительно ниже (35–56 %). Для повышения надежности метода был предложен и успешно апробирован алгоритм постобработки, отфильтровывающий ложные ноты по порогу длительности.

Abstract. The article discusses the development and experimental evaluation of a novel method for creating digital audio fingerprints based on Automatic Music Transcription (AMT) to enhance the efficiency of pirated content detection. The methodological foundation is an algorithm that converts an audio signal into a symbolic representation (MIDI) using a specially trained neural network model, "Google Onsets and Frames". This model, whose architecture includes two U-Net blocks and BiLSTM layers, was trained on the MAPS dataset of piano recordings. The Levenshtein distance was used to assess the similarity between the original and predicted MIDI representations. Experimental results revealed a strong dependence of the model's accuracy on the nature of the audio data. The highest accuracy (89.36%) was achieved on polyphonic piano compositions that matched the training data, while the results were significantly lower (35-56%) for monophonic and guitar tracks. To improve the method's reliability, a post-processing algorithm filtering out false notes based on a duration threshold was proposed and successfully tested. This procedure enabled achieving 100% similarity for most test piano recordings, proving the approach's effectiveness. Thus, the study demonstrates

Эта процедура позволила достичь 100% схожести для большинства тестовых фортепианных записей, что доказывает эффективность подхода. Таким образом, исследование демонстрирует перспективность использования AMT для создания устойчивых цифровых отпечатков и пути для дальнейшего улучшения модели, включая расширение обучающей выборки данными различных инструментов.

Ключевые слова: цифровой отпечаток, музыкальная транскрипция, машинное обучение, аудиофайлы

the promise of using AMT for creating robust digital fingerprints and outlines paths for further model improvement, including expanding the training dataset with data from various instruments.

Keywords: digital fingerprint, automatic music transcription, machine learning

Для цитирования: *Ладыгин П.С., Неверов А.И.* Методика оценки схожести аудиофайлов с применением AMT-модели // Проблемы правовой и технической защиты информации. 2025 № 12. С. 23–29.

For citation: *Ladygin P.S., Neverov A.I.* A Method for Assessing Audio File Similarity Using an AMT Model. *Problems of legal and technical protection of information*. 2025. No. 13. P. 23–29.

С переходом от физических носителей к потреблению музыки в электронном формате наблюдается непрерывное развитие новых способов получения доступа к музыкальному контенту в цифровой среде. Интернет-площадки, медиа-платформы и социальные сети продолжают улучшать и развивать сервис, позволяя получить более качественный пользовательский опыт для потребителей контента, в том числе музыкального. Такие отечественные платформы как «Яндекс Музыка», «VK Музыка» и «Звук» позволяют пользователям абсолютно правомерно приобрести доступ к прослушиванию различных музыкальных произведений за определенное вознаграждение, заключая лицензионные соглашения с правообладателями на размещение музыкальных произведений на своих площадках и выплачивая правообладателям получаемое от пользователей вознаграждение пропорционально количеству прослушиваний [1].

Такой механизм распространения музыкальных произведений является наиболее эффективным, поскольку соответствует интересам как правообладателей, так и общества. Однако, как показывает практика, можно найти сотни копий аудиофайлов на различных web-страницах, размещенных неправомерно. Кроме того, названия музыкальных файлов при несанкционированном распространении часто меняются, что затрудняет их выявление правообладателями. Поэтому для защиты своих прав автору не только необходимо отыскать все копии, существующие в сети, но и доказать первоочередность своего авторства [2].

Технические средства защиты интеллектуальной собственности предотвращают либо ограничивают осуществление действий, которые не разрешены автором или иным правообладателем в отношении произведения [3]. Наиболее распространены технологии цифровых отпечатков.

Технологии формирования звуковых отпечатков уже используются для детектирования музыкальных композиций на различных платформах, а также при добавлении аудио и видеофайлов на популярный видеохостинг «Youtube» [4]. Однако у «пиратов» получается обходить подобные алгоритмы, нарушая целостность файлов

при помощи замедления или ускорения, повышения или понижения высоты тона, а также при помощи внедрения различного рода искажений, приводящих к зашумлению медиконтента. После подобных аугментаций вычисляемые алгоритмом отпечатки не совпадают с существующими в базе данных, тем самым «пират» может загрузить чужую интеллектуальную собственность от своего имени и монетизировать ее. Ввиду этого существует необходимость в разработке новых технических методов оценки степени схожести аудиофайлов, которые могут быть использованы отечественными платформами для урегулирования неправомерного использования чужого интеллектуального труда и нарушения авторских прав в сети Интернет.

Одним из существующих подходов к решению данной задачи является использование машинного обучения для распознавания мелодических конструкций, содержащихся в аудиофайлах, поскольку именно они являются важнейшей ценностью в судебных разбирательствах по выявлению фактов плагиата. Для выявления вектора признаков из спектрограммы аудиофайла, для формирования цифрового отпечатка может быть использована модель автоматической музыкальной транскрипции.

Автоматическая музыкальная транскрипция (AMT) — является фундаментальной проблемой считывания музыкальной информации из аудиофайлов. AMT нацелена на извлечение музыкальной информации из аудиофайлов и ее записи в символическом представлении в виде различного рода музыкальных нотаций [5].

Одним из основных способов хранения музыкальной нотации в цифровом виде является стандарт MIDI (Musical Instrument Digital Interface), который представляет собой наборы команд (проигрываемые ноты, значения изменяемых параметров звука, время и сила нажатия на клавишу). Высота каждой ноты в мидифайле кодируется числом, например, ноте «До в большой октаве» соответствует число 48. Таким образом мидифайл можно представить в виде таблицы, содержащей в себе музыкальные параметры.

Основываясь на данном представлении, а также на возможности извлечения музыкальной информации и получения нотации при помощи AMT-модели, предлагается следующий алгоритм получения цифрового отпечатка из аудиофайла (рисунок 1). Аудиофайл подается на вход специально обученной модели автоматической музыкальной транскрипции, целью которой является извлечение музыкальной информации из аудиофайла и запись данной информации в мидифайл, из которого программа вычленяет параметры, отвечающие за начало и конец ноты, а также за высоту самой ноты.

Рисунок 1. Блок-схема получения цифрового отпечатка аудиофайла

Для обучения нейросети была выбрана модель с открытым исходным кодом «Google Onsets and Frames», адаптированная под библиотеку «pytorch». Данная модель относится ко второму типу, поскольку она может транскрибировать несколько временных шагов одновременно в зависимости от размера входной спектрограммы. Архитектура рассматриваемой модели представлена на рисунке 2.

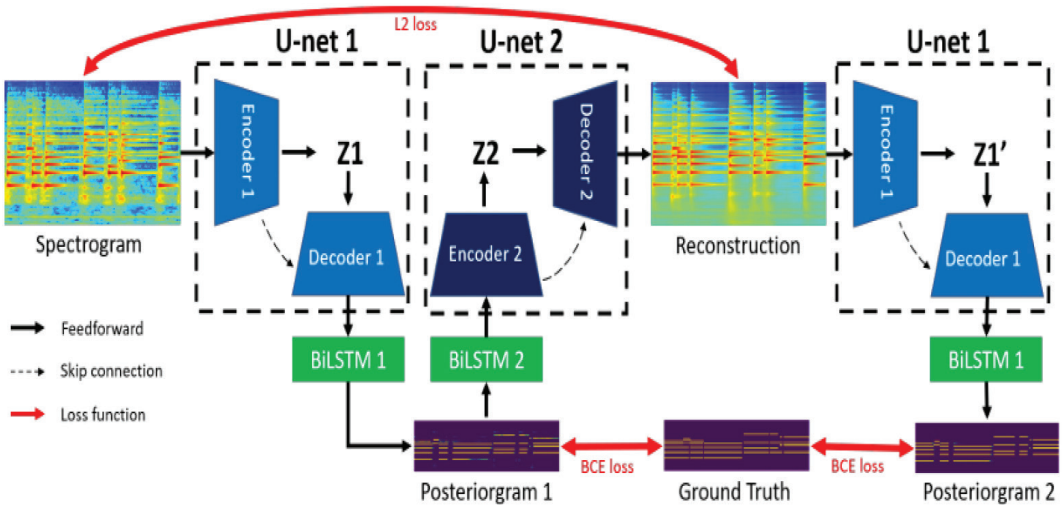

Рисунок 2. Архитектура модели «Google onsets and frames» [5]

Рассматриваемая модель состоит из двух U-NET блоков. U-net 1 и U-net 2 имеют идентичную архитектуру: четыре блока свертки для части кодировщика и четыре блока свертки для части декодера. Двунаправленные LSTM (BiLSTM) расположены после U-net 1 и перед U-net 2 для моделирования меж кадровых зависимостей. BiLSTM после U-net 1 преобразует выходные данные U-net 1 в постериограммы. Эти постериограммы затем используются для реконструкции спектрограмм через еще один BiLSTM и U-net 2. Внутри каждого U-net есть только пропускные соединения, другими словами, U-net 2 обязан реконструировать спектрограммы, используя информацию, доступную из постериограмм.

Мысль, лежащая в основе данной архитектуры, заключается в том, что, люди транскрибируя музыку, обычно слушают, как звучит их транскрипция и сравнивают ее с оригинальной аудиозаписью, проверяя правильность [6]. Реконструктор «U-net 2» пытается смоделировать этот процесс. Реконструированная спектрограмма является очищенной версией оригинальной спектрограммы, которая впоследствии снова передается на блок «U-net 1» для окончательной транскрипции.

В области оценки автоматической музыкальной транскрипции (AMT), ощущается потребность в качественных наборах данных. Они необходимы как для разработки, так и для оценки алгоритмов, однако при создании подобных баз данных возникает ряд проблем. Среди которых главную роль занимают: сложность производства и зависимость от авторского права.

В качестве обучающего набора данных был выбран «MAPS» (MIDI Aligned Piano Sounds [7]). «MAPS» — это база данных, состоящая из аудиофайлов фортепианных произведений с аннотациями в MIDI формате. «MAPS» предоставляет собой записи с CD качеством (стереозвук с частотой дискретизации 44,1 кГц) и соответствующие согласованные MIDI-файлы в качестве аннотации.

Для обучения модели использовалось 210 аудиофайлов из «MAPS», для тестирования модели использовалось 60 аудиофайлов на основе реального фортепиано, записанного самостоятельно. Обучение модели проводилось в 1000 итераций, при этом тестирование модели проводилось каждые 100 эпох в автоматическом режиме для отслеживания улучшения точности модели. Так же каждые 100 итераций в автоматическом режиме про-

водилось отслеживание изменения показателей точности предсказания времени и высоты ноты от количества пройденных эпох.

Для тестирования модели в ручном режиме были созданы 5 аудиофайлов, на основе самостоятельно записанных сэмплов фортепиано и акустической гитары на основе цифровой рабочей станции «YAMAHA PSR-SX700», с вручную созданными аннотациями.

Результаты оценки модели представлены в таблице 1. В качестве оценки степени схожести использовалось расстояние Левенштейна [8] в процентном соотношении между оригинальным MIDI-файлом с вычисленным с помощью AMT.

Таблица 1. Результаты оценки работы модели на самостоятельно созданных аудиофайлах

Описание аудиофайла	Степень схожести, %
Полифонический аудиофайл, составленный из сэмплов звуков фортепиано, в тональности «до мажор».	89,36
Полифонический аудиофайл, составленный из сэмплов звуков акустической гитары, в тональности «до мажор».	76,36
Монофонический аудиофайл, составленный из сэмплов звуков фортепиано, представляющий собой гамму «до мажор».	56,00
Монофонический аудиофайл, составленный из сэмплов звуков фортепиано, представляющий собой первый такт произведения «в траве сидел кузнечик».	41,18
Монофонический аудиофайл, составленный из сэмплов звуков акустической гитары, представляющий собой гамму «до мажор».	35,00

Исходя из полученных результатов можно сделать следующие выводы относительно работы модели:

1) так как модель обучалась на фортепианных аудиофайлах, большая часть которых полифонические, модель лучше всего показывает себя именно на таких файлах;

2) модель успешно справляется с гармонической последовательностью, составленной из сэмплов акустической гитары, уступая результатам предсказания фортепианных звуков на 13% на рассматриваемых аудиофайлах;

3) модель сравнительно хуже показывает себя в предсказании монофонических произведений, что объясняется выбором обучающего набора данных;

4) хуже всего модель показывает себя на монофоническом аудиофайле, состоящем из сэмплов акустической гитары, то есть на звуках, с которыми не сталкивалась в процессе обучения.

MIDI-файлы, представляющие собой исходную аннотацию и предсказание модели для третьего аудиофайла (см. таблицу 1) представлены на рисунках 3 и 4 соответственно. На рисунке 3 видно, как модель пытается распознать многоголосье там, где его нет и предсказывает несуществующие ноты, которые отмечены красным цветом. Такое поведение модели объясняется тем, что обучающий набор данных состоял преимущественно из полифонических звуков.

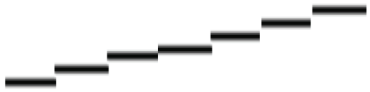

Рисунок 3. Исходная аннотация

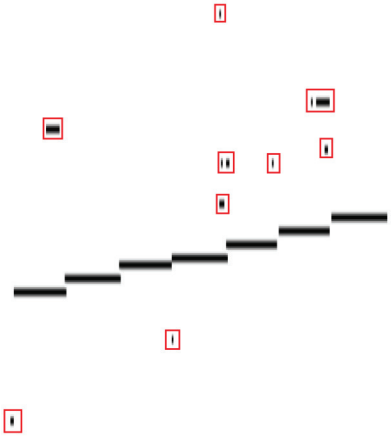

Рисунок 4. Результат работы модели

Корректная работа предложенного метода происходит в случае, если аудиофайл содержит одну мелодическую линию, сыгранную на фортепиано, где в каждый момент времени звучит одна нота. В таких случаях AMT будет достаточно точно распознавать мелодию, однако с целью увеличения точности работы модели возможна реализация «отсеивания» ложных предсказаний. В рассматриваемом на рисунке 3 аудиофайле длительность верно предсказанных нот составляет 0,775 секунды, тогда как длительность большинства ложно предсказанных нот составляет примерно 0,05 секунды. Исходя из этого можно сделать вывод что эффективным способом фильтрации будет установление порога для значения длительности ноты. Результаты оценки модели после проведения предварительной фильтрации представлены в таблице 2.

Таблица 2. Результаты оценки работы модели после предварительной фильтрации

Описание аудиофайла	Степень схожести до фильтрации, %	Степень схожести после фильтрации, %	Увеличение показателя степени схожести, %
Полифонический аудиофайл, составленный из сэмплов звуков фортепиано, в тональности «до мажор».	89,36	100,00	10,64
Полифонический аудиофайл, составленный из сэмплов звуков акустической гитары.	76,36	100,00	23,64
Монофонический аудиофайл, составленный из сэмплов звуков фортепиано, представляющий собой гамму «до мажор».	56,00	100,00	44,00
Монофонический аудиофайл, составленный из сэмплов звуков фортепиано, представляющий собой первый так произведения «в траве сидел кузнечик».	41,18	100,00	58,82
Монофонический аудиофайл, составленный из сэмплов звуков акустической гитары.	35,00	40,00	5,00

Значение порога было установлено в 0,5 секунды что соответствует «шестнадцатой длительности» в темпе 60 bpm, все ноты, чья длительность ниже данного порога считались неверно предсказанными. Как видно, установление такого порога для проведения фильтрации ложно предсказанных нот позитивно сказывается на оценке степени схожести предсказаний модели с исходными аннотациями.

В заключение можно констатировать, что разработанный метод формирования цифрового отпечатка на основе AMT-модели и последующей фильтрации предсказаний показал свою высокую эффективность. Предложенное решение позволяет преодолеть одно из ключевых ограничений подхода — генерацию ложных нот при транскрипции несвойственных модели данных, что подтверждается значительным ростом метрики схожести после применения порога по длительности ноты. Таким образом, несмотря на существующие вызовы, связанные с обобщающей способностью модели, исследование закладывает основу для создания более надежных и адаптивных систем аудиоидентификации, способных противостоять современным методам обхода защиты.

Библиографический список

1. Чевтаева Л.Н. Интернет-пиратство: вчера и сегодня // Вестник Саратовского государственного технического университета. 2013. С. 1–6.

2. Борисова С.Н. Методы защиты аудиофайлов от несанкционированного копирования и распространения // Фундаментальные исследования. 2015. № 5 (часть 3). С. 481–487.

3. Гражданский кодекс Российской Федерации (часть четвертая) от 18.12.2006 N 230-ФЗ (ред. от 03.07.2016, с изм. от 13.12.2016) (с изм. и доп., вступ. в силу с 01.01.2017) // КонсультантПлюс: справочно-правовая система. URL: http://www.consultant.ru/document/cons_doc_LAW_64629/ (дата обращения: 5.06.2025).

4. Google LLC: Как работает система Content ID : сайт. 2025. URL: https://support.google.com/youtube/answer/2797370?hl=ru (дата обращения: 24.04.2025).

5. Андрадэ А.И., Насуро Е.В. Средство музыкальной транскрипции при помощи методов машинного обучения // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сборник материалов V Международной научно-практической конференции, Минск, 13–14 марта 2019 г.: в 2 ч. Ч. 1 / Белорусский государственный университет информатики и радиоэлектроники; редкол. : В.А. Богуш [и др.]. Минск, 2019. С. 376–380.

6. Kin Wai Cheuk ReconVAT: A Semi Supervised Automatic Music Transcription Framework for Low-Resource Real-World Data / Kin Wai Cheuk, Dorien Herremans, Li Su // ACM International Conference on Multimedia (China, 24.10.2021), China, 2021. С. 1–9.

7. ADASP: MAPS Database: a Piano database for multipitch estimation and automatic transcription of music : сайт. 2025. URL: https://adasp.telecom-paris.fr/resources/2010-07-08-maps-database/ (дата обращения: 13.02.2025).

8. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады АН СССР, 1965. Т. 163. № 4. С. 845–848.