Automatic Music Transcription

A Preprint

Дмитрий Протасов
МФТИ
dmitry.protasov@gmail.com

Консультант: Дмитрий Ковалев

Эксперт: д.ф.-м.н. Иван Матвеев

МФТИ, 2023

Abstract

Автоматическая транскрипция музыки (AMT) остается важной, но сложной задачей в области поиска музыкальной информации, которую затрудняют ограниченные наборы данных MIDI и низкое качество существующих моделей. Данное исследование направлено на повышение точности транскрипции за счет использования специализированных моделей для извлечения различных музыкальных особенностей, таких как аккордовые прогрессии, тональность, ритм и типы инструментов. Для решения проблемы нехватки наборов MIDI-данных мы предлагаем использовать синтетические данные для пополнения обучающих ресурсов. Этот подход предлагает новый способ потенциального обогащения моделей AMT и развития этой области.

1 Introduction

Автоматическая транскрипция музыки (AMT) - ключевая задача в области извлечения информации из музыки, которая заключается в преобразовании аудиосигналов в символическое представление. Исследования в области AMT мотивированы широким спектром ее применения - от помощи в музыковедческом анализе до облегчения музыкального образования и производства. Цель этого исследования – усовершенствование процесса транскрипции путём включения анализа тональности и определения количества ударов в минуту (BPM) для повышения точности идентификации нот

AMT - сложная задача, обусловленная сложной природой полифонического звука, в котором несколько инструментов и голосов накладываются друг на друга. MT3 (Gardner [2022]) представил подход к транскрипции для множества инструментов с использованием модели трансформера, установив новый эталонный уровень для данной области. Однако этот метод в значительной степени зависит от больших нейронных сетей, требующих существенных вычислительных ресурсов. Работа Jointist (Cheuk [2023]) разделяет проблему на три подзадачи: разделение музыкальных источников, распознавание инструментов и саму транскрипцию. Хотя такое разделение предлагает структурированный подход, отсутствие общедоступного кода ограничивает его воспроизводимость и дальнейшее развитие.

The novel contribution of basic-pitch (Bittner [2022]) is the use of Constant-Q Transform (CQT) for monophonic transcription, which aligns more naturally with musical theory compared to mel-spectrograms. However, its limitation to monophonic audio restricts its applicability to more complex compositions. Furthermore, a comprehensive review by (Solovyev [2023]) evaluates the current landscape of music source separation, which is a foundational step in AMT.

A recent study by (Sato [2023]) introduces a synthetic dataset for AMT; however, this approach may lead to impoverished sound representations due to its synthetic nature. In contrast, this research aims to leverage real-world datasets to capture the rich nuances present in actual music recordings.

Чтобы устранить недостатки существующих моделей, в данной работе предлагается гибридное решение, использующее дискретные элементы AMT, такие как определение клавиш и оценка BPM, предполагая, что эти музыкальные аспекты могут направлять и уточнять процесс транскрипции нот. 4

2 Problem Statement

Автоматическая транскрипция музыки (AMT) направлена на преобразование музыкальных аудиосигналов в символическое представление, в частности в формат MIDI, в котором подробно описываются музыкальные ноты, их время и динамика. Это предполагает идентификацию и выделение отдельных музыкальных нот и инструментов из сложного аудиосигнала и точную транскрипцию этой информации в структурированный цифровой формат, который может быть использован для решения различных музыковедческих и вычислительных музыкальных задач.

Рассмотрим аудиосигнал $A(t) : [0, T] \to \mathbb{R}$, где T – продолжительность сигнала в секундах, представляющий музыкальное произведение по времени t. Здесь t обозначает дискретное время в рамках частоты дискретизации, обычно равной 22050 Гц, что означает, что каждую секунду записи представляет 22050 сэмплов. Значения сигнала $A(t)$ могут быть представлены, например, как 16-битные целые числа (int16), охватывая диапазон от -32768 до 32767, где каждое значение соответствует мгновенной амплитуде звукового сигнала в этот момент времени. Цель автоматической музыкальной транскрипции (AMT) – транскрибировать этот аудиосигнал в последовательность событий MIDI, которые захватывают музыкальное содержание, включая начала нот, окончания и высоты тона для каждого инструментального трека, исключая динамику и ударные для упрощения.

Пусть $S = \{(n_i, t_{on_i}, t_{off_i}) | i = 1, \ldots, N\}$ будет целевой последовательностью событий MIDI для данного инструмента, где n_i представляет номер MIDI ноты, t_{on_i} и t_{off_i} - время начала и окончания i-й ноты, а N – общее количество нот.

Наша модель, M, отображает аудиосигнал в предсказанную последовательность событий MIDI: $M : A(t) \to S'$. Цель состоит в том, чтобы найти параметры модели, минимизирующие расхождение между предсказанной последовательностью S' и целевой последовательностью S. 5

Given the discrete nature of MIDI events, we employ a cross-entropy loss function for optimization. For each time frame j and each possible note n, we define a probability distribution over the possible states (note on, note off) predicted by the model. The cross-entropy loss L for a single note event is then given by:

$$L = -\sum_{j=1}^{J} \sum_{n=1}^{128} y_{jn} \log(p_{jn}) + (1 - y_{jn}) \log(1 - p_{jn}), \tag{1}$$

where J is the total number of time frames, y_{jn} is the binary indicator (0 or 1) of the presence of note n in time frame j in the target sequence, and p_{jn} is the predicted probability of note n's presence in time frame j.

The optimization problem can thus be formulated as:

$$\underset{M}{\operatorname{argmin}} \sum_{i=1}^{I} L(M(A_i(t)), S_i), \tag{2}$$

6

where I is the number of instances (audio tracks) in the dataset.

Для оценки качества транскрибированных музыкальных записей, наряду с минимизацией функции потерь, используется метрика 7 F_{no}, которая учитывает точность, полноту и перекрытие временных интервалов нот. Метрика вычисляется на основе сравнения между эталонными (reference) и оценочными (estimated) нотами. Нота считается правильно транскрибированной, если её начало находится в пределах ±50 мс от начала эталонной ноты, высота тона в пределах ±50 центов от соответствующей эталонной ноты, а окончание ноты – в пределах 20% (по умолчанию) от длительности эталонной ноты вокруг окончания эталонной ноты или в пределах не менее 50 мс, в зависимости от того, что больше. Если 8 параметр «offset_ratio» установлен в значение «None», окончания нот не учитываются при сравнении.

Пусть t_{on}^{ref} и t_{on}^{est} обозначают времена начала эталонной и оценочной нот соответственно, f^{ref} и f^{est} – их высоты тонов в центах, а t_{off}^{ref} и t_{off}^{est} – времена окончания. Тогда условия можно выразить следующим образом:

1. **Условие начала ноты:**

$$|t_{on}^{est} - t_{on}^{ref}| \leq 50 \text{ мс}$$

2. **Условие высоты тона:**

$$|f^{est} - f^{ref}| \leq 50 \text{ центов}$$

3. **Условие окончания ноты**: Определим длительность эталонной ноты как $d^{ref} = t_{off}^{ref} - t_{on}^{ref}$ и установим порог окончания ноты, Δt_{off}, как максимум из 20% длительности эталонной ноты и 50 мс:

$$\Delta t_{off} = \max(0.2 \cdot d^{ref}, 50 \text{ мс})$$

Тогда условие для окончания ноты выглядит так: $|t_{off}^{est} - t_{off}^{ref}| \leq \Delta t_{off}$

Если «offset_ratio» установлен в значение «None», условие окончания ноты игнорируется.

Точность (*Precision*), полнота (*Recall*) и F-мера ($F_{measure}$) вычисляются следующим образом:

$$Precision = \frac{\text{Number of correctly transcribed notes}}{\text{Total number of estimated notes}},$$

$$Recall = \frac{\text{Number of correctly transcribed notes}}{\text{Total number of reference notes}},$$

$$F_{measure} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

Дополнительно вычисляется Средний Коэффициент Перекрытия (Average Overlap Ratio), который оценивает среднее перекрытие временных интервалов между правильно транскрибированными оценочными и эталонными нотами.

3 Планирование эксперимента

Для решения недостатков существующих моделей, в данной работате предлагается гибридное решение, которое использует дискретные элементы AMT, такие как определение тональности (key detection) и оценка количества ударов в минуту (BPM estimation), предполагая, что эти музыкальные аспекты могут направлять и уточнять процесс транскрипции нот.

Эксперименты были спроектированы для проверки этой гипотезы с использованием набора данных BabySlakh, который является устоявшимся ресурсом в сфере автоматической музыкальной транскрипции. Этот датасет содержит мульти-трековые MIDI-транскрипции, позволяя детально аннотировать взаимодействие между тональностью, ритмом и событиями нот

В следующих разделах будет подробно изложена методология, включая обзор литературы и детальное изучение передовых техник. Будут очерчены задачи проекта, а также подчеркнута новизна и преимущества предложенного решения по сравнению с современными моделями.

3.1 Теоретическое обоснование улучшения с помощью BPM и тональности

Предложенное улучшение базируется на теории музыки, согласно которой знание BPM и тональности композиции позволяет более точно предсказывать начало и окончание нот, а также их принадлежность к музыкальной шкале. Рассмотрим следующие аспекты:

Квантизация BPM. Если известен BPM композиции, можно предположить, что начало и окончание каждой ноты соответствуют целочисленному количеству ударов в рамках музыкального такта. Это

позволяет квантизировать временные рамки нот, значительно улучшая точность определения их временных интервалов. Математически это можно выразить как:

$$t_{on}^{quant} = \left\lfloor \frac{t_{on} \cdot \text{BPM}}{60} \right\rceil \cdot \frac{60}{\text{BPM}}, \quad t_{off}^{quant} = \left\lfloor \frac{t_{off} \cdot \text{BPM}}{60} \right\rceil \cdot \frac{60}{\text{BPM}},$$

где t_{on}^{quant} и t_{off}^{quant} – квантизированные времена начала и окончания ноты, t_{on} и t_{off} – исходные времена начала и окончания, а BPM – темп композиции в ударах в минуту.

Фильтрация по тональности. Знание тональности позволяет отфильтровать ноты, которые не соответствуют музыкальной гармонии композиции. Для каждой ноты, обозначаемой как n с высотой звука в формате, например, C4 (где буква обозначает ноту, а число – октаву), можно определить, соответствует ли она текущей тональности композиции. Если композиция находится в тональности C мажор, ноты, не входящие в шкалу C мажора, могут быть исключены из предсказания. Определим функцию фильтрации $F_{\text{key}} : \mathcal{N} \to \{0, 1\}$, где $\mathcal{N}$ – пространство всех возможных нот. Функция возвращает 1 для нот n, принадлежащих заданной тональности, и 0 — в противном случае. Математически это можно представить как:

$$F_{\text{key}}(n) = \begin{cases} 1 & \text{если } n \in \text{Тональность}, \\ 0 & \text{иначе.} \end{cases}$$

15

Этот подход позволяет эффективно сократить количество потенциальных ошибок в транскрипции, исключая ноты, которые маловероятно будут играть в данной тональности, тем самым улучшая общую точность идентификации нот.

Эти методы позволяют существенно повысить точность автоматической музыкальной транскрипции, опираясь на фундаментальные принципы музыкальной теории и структуры.

4 Эксперимент и результаты

Целью вычислительного эксперимента является проверка гипотезы о том, что интеграция анализа тональности и квантизации BPM в процесс автоматической музыкальной транскрипции может значительно улучшить точность идентификации нот

4.1 Постановка и условия эксперимента

Экспериментальная оценка проводилась на датасете BabySlakh, который представляет собой обширный набор MIDI-транскрипций, сгенерированных из мульти-трековых аудиозаписей. Для оценки эффективности предложенного подхода использовались метрики F_{no} с различными параметрами «offset», чтобы изучить влияние учёта длительности нот на качество транскрипции.

4.2 Описание алгоритма

Эксперимент включал в себя сравнение четырёх конфигураций: 1. Базовая модель AMT без дополнительных уточнений. 2. Модель AMT с интеграцией маскирования на тональность (key masking). 3. Модель AMT с квантизацией BPM (bpm quantization). 4. Модель AMT с одновременным использованием анализа тональности и квантизацией BPM.

4.3 Результаты

Для демонстрации результатов использовались следующие метрики: F_{no} без учета «offset» и F_{no} с параметром «offset», равным 0.2.

Модель	$F_{no}^{offset=0.2}$	$F_{no}^{offset=None}$
Обычная модель	TODO	TODO
+ Key Masking	TODO	TODO
+ BPM Quantization	TODO	TODO
+ Key + BPM	TODO	TODO

Таблица 1: Сравнение эффективности различных конфигураций модели на датасете BabySlakh

4.4 Анализ результатов

TODO

16, 17

Список литературы

Rachel M. Bittner. Basic pitch: A basic model for pitch detection in polyphonic music. In *Conference Name*, 2022.

Kin Wai Cheuk. Jointist: A multi-faceted approach to music-source-separation, instrument-recognition, and transcription. *https://arxiv.org/abs/2302.00286*, Volume(Number):Pages, 2023.

Josh Gardner. Mt3: Multi-task multitrack music transcription. *arXiv preprint arXiv:2111.03017*, 2022.

Gakusei Sato. A synthetic dataset for automatic music transcription. *arXiv preprint arXiv:2312.10402*, 2023.

R. Solovyev. A comprehensive review of music source separation. *arXiv preprint arXiv:2305.07489*, 2023.

1. Здесь и по тексту дальше - здесь нужно использовать длинное тире.
В латехе набирается как три дефиса ("---")

2. Нужна какая-то фундаментальная/обзорная работа, которая показывает что задача есть и она серьезная

3. На каждыйы пример (помощь в анализе, облегчение образования) нужна ссылка на статью, подтверждающую этот пример

4. Здесь неплохо было бы дописать на чем, для чего и как проводится эксперимент

5. Формулы нужно стараться не разрывать

6. по какому множеству идет оптимизация?

7. лучше не метрика, а критерий качества. Метрика - термин из анализа.

8. Вот эта часть про offset_ratio - чересчур техничная.
Как вариант - ввести два варианта метрики со своими буквами или индексами и объяснить различие

9. После формул идут знаки препинания

10. Здесь как-то внезапно возник дефис

11. Чаще называют F-score (и это более корректно, чтобы не смешивать термин с мерой из анализа)

12. С этой секцией нужно сильно поработать, пока структурировано слабо. Плюс куча текста на английском, мне неясно, был ли он откуда-то скопирован или нет.

13. Лучше объяснить AOR формулой

14. Нужно совместить с экспериментом

15. Так не пишут. Введите букву для множества, отвечающего тональности.

16. Нужен анализ результатов и заключение.

17. Список литературы очень "свежий".
Добавьте более классических статей в обзор, иначе появляется ощущение что задача родилась вчера.