Auotmatic Music Transcription

Дмитрий Протасов

МФТИ, 2024

9 апреля 2024 г.

One slide talk

Figure 1. The proposed Jointist framework. Our actual framework can transcribe/separate up to 39 different instruments as defined in Table 7 of Appendix. B: batch size, L: audio length, C: instrument classes, T: number of time steps, K: number of predicted instruments. Dotted lines represent iterative operations for K times. Best viewed in color.

Goal:

$$\text{Audio} \rightarrow \text{MIDI (events note, on/off, time)}$$

Method:

- Use *Demucs* for extracting vocal and instrumental parts.
- Apply *Basic Pitch*, *Key Detection*, and *BPM Estimation* to enhance vocal transcription.

Literature

- MT3: Multi-Task Multitrack Music Transcription, 2022.
- Jointist: A multi-faceted approach to music-source-separation, instrument-recognition, and transcription, 2023.
- Basic-pitch: lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation, 2022.

Problem Statement

Автоматическая музыкальная транскрипция (AMT) направлена на преобразование аудиосигналов в символическое представление. Рассматривая аудиосигнал $A(t) : [0, T] \to \mathbb{R}$, целью AMT является транскрибировать его в последовательность событий MIDI $S = \{(n_i, t_{on_i}, t_{off_i}) | i = 1, \ldots, N\}$, где n_i - номер MIDI ноты, t_{on_i} и t_{off_i} - время начала и окончания ноты.

Оптимизация задачи описывается как:

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

где L - функция потерь, рассчитанная через CrossEntropyLoss для MIDI событий, $\{(A_i(t), S_i) | i = 1, ..., M\}$ – обучающая выборка

Problem Solution

Предлагается гибридное решение, интегрирующее элементы AMT, такие как определение тональности и оценка BPM. Теоретическое обоснование улучшения включает:

Квантизация BPM:

$$t_{on}^{quant} = \left\lfloor \frac{t_{on} \cdot BPM}{60} \right\rceil \cdot \frac{60}{BPM}, \tag{2}$$

$$t_{off}^{quant} = \left\lfloor \frac{t_{off} \cdot BPM}{60} \right\rceil \cdot \frac{60}{BPM}, \tag{3}$$

Фильтрация по тональности:

$$F_{key}(n) = \begin{cases} 1, & \text{если } n \in \text{Тональность} \\ 0, & \text{иначе} \end{cases} \tag{4}$$

Это позволяет повысить точность транскрипции, сокращая количество потенциальных ошибок.

Метрика F_{no}

Метрика F_{no} используется для оценки качества транскрибированных музыкальных записей, учитывая точность, полноту и перекрытие временных интервалов нот. Нота считается правильно транскрибированной, если выполняются следующие условия:

- Начало ноты находится в пределах ± 50 мс.
- Высота тона в пределах ± 50 центов.
- Если параметр «offset_ratio» не равен «None», окончание ноты должно находиться в пределах 20% от длительности эталонной ноты или не менее 50 мс, в зависимости от того, что больше.

Формулы для расчета:

$$Precision = \frac{\text{Количество правильно транскрибированных нот}}{\text{Общее количество оценочных нот}},$$

$$Recall = \frac{\text{Количество правильно транскрибированных нот}}{\text{Общее количество эталонных нот}},$$

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

Computational Experiment

Эксперименты проводились на датасете BabySlakh с целью проверки гипотезы об улучшении точности транскрипции через анализ тональности и BPM. Использовались метрики F_{no} с разными параметрами «offset» для оценки эффективности подходов, включая базовую модель AMT, модель с анализом тональности, модель с квантизацией BPM, и комбинированную модель.

Results and Analysis

Результаты показали, что добавление анализа тональности и квантизации BPM улучшает точность транскрипции по сравнению с базовой моделью, при этом наибольшее улучшение наблюдается при их совместном использовании. Это подтверждает предположение о важности учета музыкальной структуры и ритмической сетки в процессе AMT.

Модель	F_{no} (без «offset»)	F_{no} (с «offset»)
Обычная модель	0.65	0.60
+ Key Estimation	0.68	0.63
+ BPM Quantization	0.70	0.65
+ Key + BPM	0.75	0.70

Таблица: Сравнение эффективности различных конфигураций модели на датасете BabySlakh.