Automatic Music Transcription

Dmitry Protasov

MIPT, 2023

Abstract

Automatic music transcription (AMT) remains an important but challenging task in music information retrieval, hampered by limited MIDI datasets and the poor quality of existing models. This research aims to improve transcription accuracy by using specialized models to extract distinct musical features such as chord progressions, tonality, rhythm, and instrument types. To address the scarcity of MIDI datasets, we propose the use of synthetic data to augment training resources. This approach offers a new way to potentially enrich AMT models and advance the field.

1 Introduction

Automatic Music Transcription (AMT) is a pivotal task in music information retrieval that involves converting audio signals into a symbolic representation. The research in AMT is motivated by its vast array of applications, from aiding musicological analysis to facilitating music education and production. The objective of this research is to refine the transcription process by incorporating key detection and beats per minute (BPM) analysis to improve the accuracy of note identification, particularly focusing on vocal parts prevalent in polyphonic music.

AMT is a complex challenge due to the intricate nature of polyphonic audio, where multiple instruments and voices overlap. MT3 introduced a multi-instrument transcription approach using a transformer model, setting a new state-of-the-art benchmark. However, this method relies heavily on large neural networks, which require substantial computational resources. The work of Jointist dissects the problem into three sub-tasks: music source separation, instrument recognition, and transcription itself. While this dissection provides a structured approach, the lack of publicly available code limits its reproducibility and further development.

The novel contribution of basic-pitch 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 [**?**] evaluates the current landscape of music source separation, which is a foundational step in AMT.

A recent study by [] 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.

To address the deficiencies of current models, this project proposes a hybrid solution that employs discrete AMT elements such as key detection and BPM estimation, hypothesizing that these musical aspects can guide and refine the note transcription process. By concentrating predominantly on vocal parts, the research taps into a universal element present across various genres, thereby ensuring a wide applicability of the findings.

The experiments are designed to validate this hypothesis using well-established datasets and a workflow that meticulously annotates the interplay between key, rhythm, and note events. By doing so, this research not only contributes to the body of knowledge in AMT but also propels the practical utility of transcription systems.

In the following sections, the methodology, including a literature review and a detailed examination of state-of-the-art techniques, will be expounded. The project tasks will be outlined, and the proposed solution's novelty and advantages over recent models will be underscored.

2 Problem Statement

Automatic Music Transcription (AMT) seeks to convert audio signals of music into a symbolic representation, specifically MIDI format, which details the musical notes, their timing, and dynamics. This involves identifying and isolating individual musical notes and instruments from complex audio inputs and accurately transcribing this information into a structured digital format that can be used for various musicological and computational music tasks.

Consider an audio signal $A(t)$ representing a musical piece over time t. The aim of AMT is to transcribe this audio signal into a sequence of MIDI events that capture the musical content, including note onsets, offsets, and pitches, for each instrument track, while excluding dynamics and percussion for simplicity.

Let $S = \{(n_i, t_{on_i}, t_{off_i}) | i = 1, \ldots, N\}$ be the target sequence of MIDI events for a given instrument, where n_i represents the MIDI note number, t_{on_i} and t_{off_i} are the onset and offset times of the i-th note, and N is the total number of notes.

Our model, M, maps the audio signal to a predicted sequence of MIDI events: $M : A(t) \rightarrow S'$. The goal is to find the model parameters that minimize the discrepancy between the predicted sequence S' and the target sequence S.

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:

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

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