Automatic Music Transcription

Dmitry Protasov

MIPT, 2023

Abstract

This article discusses the problem of automatic music transcription, which involves converting audio representations of songs into their MIDI representations. The goal is to explore and improve existing algorithms for extracting MIDI from songs. The article covers various aspects of this problem, including Music Source Separation, Instrument Recognition, and Note Transcription. It also reviews relevant literature and discusses metrics used to evaluate transcription quality.

1 Introduction

Music transcription is the process of converting audio recordings of music into symbolic representations, such as MIDI (Musical Instrument Digital Interface) files. This task is essential for various applications, including music analysis, synthesis, and composition. However, most of the music available on the internet is in audio format, making it challenging to work with generative music models that require MIDI data. This article addresses the problem of automatically transcribing music from audio to MIDI and aims to improve existing algorithms in this domain.

1.1 Research Goals

The main objectives of this work are as follows:

1. Build and curate a database of songs and their MIDI representations.
2. Study and evaluate existing models for audio-to-MIDI conversion.
3. Develop and test novel methods for extracting MIDI information from audio recordings.

2 Related Works

This section provides an overview of relevant literature in the field of automatic music transcription, with a focus on works related to transcription, music source separation, and instrument recognition.

2.1 Transcription

- **MT3 (2022)**: This state-of-the-art model, based on the T5 architecture, excels in multi-instrument transcription.
- **Jointist (2022)**: It employs a combination of CNN and Transformer for instrument recognition and relies on the onsets-and-frames framework for note transcription.
- **Crepe (2018)**: Crepe is designed to find the fundamental frequency in audio, which can be useful for extracting notes from vocals.

2.2 Music Source Separation

- **Benchmarks and Leaderboards**: A comprehensive study of sound demixing tasks with benchmarks and leaderboards.
- **Demucs**: Demucs is based on a U-Net convolutional architecture and is used for music source separation.
- **MDX-Net**: MDX-Net employs a two-stream neural network for music demixing, consisting of six separately trained networks.
- **Band-split RNN**: This approach utilizes two RNNs along the frequency and time axes for music demixing.
- **MUSDB18 Dataset**: The MUSDB18 dataset comprises 150 music tracks with isolated drums, bass, vocals, and other instruments.

3 Metrics

To evaluate the quality of automatic music transcription, several metrics are commonly used. One of the key metrics is the Signal-to-Distortion Ratio (SDR). SDR measures the quality of audio separation and aims to maximize signal fidelity while minimizing distortion. It can be computed for individual stems (instruments) and for the entire recorded mix. The overall SDR is the average SDR value across multiple records in the test set.

4 Experiments

The experiments conducted in this research include the following:

1. Generation of a synthetic dataset for testing purposes.
2. Conversion of audio recordings to spectrograms.
3. Music source separation using the Demucs model.
4. Note extraction using various algorithms, including Crepe and custom implementations.

5 Conclusion

Automatic music transcription is a challenging task with various subproblems, including Music Source Separation, Instrument Recognition, and Note Transcription. This article has provided an overview of related works and discussed relevant metrics for evaluating transcription quality. Ongoing experiments aim to enhance the current algorithms and contribute to the field of music transcription.