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.

2 Related Works

In recent years, several approaches and models have been proposed for automatic music transcription. These approaches can be categorized into three main stages: Music Source Separation, Instrument Recognition, and Note Transcription. Here are some noteworthy works in each of these areas:

2.1 Music Source Separation

- **Demucs**: Demucs is a model based on a U-Net convolutional architecture that aims to separate individual sources, such as instruments, from mixed audio recordings.

- **MDX-Net**: MDX-Net is a two-stream neural network designed for music demixing. It consists of six networks, each trained separately to extract different sources from audio.

- **Band-split RNN**: This approach explores using Recurrent Neural Networks (RNNs) for music source separation.

2.2 Instrument Recognition

- **Jointist**: Jointist combines Convolutional Neural Networks (CNN) and Transformers for instrument recognition in audio. It is part of the pipeline for music transcription.

2.3 Note Transcription

- **Crepe**: Crepe is a model that estimates the pitch of monophonic audio recordings using a convolutional neural network.

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

Currently, experiments are ongoing to improve the existing algorithms for music transcription. These experiments involve testing various models, analyzing their limitations, and exploring research areas related to Music Source Separation, Instrument Recognition, and Note Transcription.

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.