MIREX 2024: Multi-Voice Transcription Task Submission

Wang Lianganzi

Queen Mary University of London

October 27, 2024

Contents

Chapter 1

Introduction

This document describes a approach to the MIREX 2024 Multi-Voice Transcription Task, which involves converting piano recordings from audio to MIDI. This method can accurately transcribes solo piano performances by extracting detailed note information, including start time, end time, pitch, and velocity, as well as sustain pedal events. To achieve this, we use two models: the Basic Pitch model for extracting note events and the High-Resolution Piano Transcription and Pedal Regression model for predicting sustain pedal events.

1.1 Submission Details

The system outputs notes with durations extended based on pedal use and includes pedal events; therefore, no additional suffixes are necessary.

1.2 System Description

This transcription system combines two models:

1.2.1 Basic Pitch [1]

- A lightweight, instrument-agnostic model for multi-voice note transcription and polyphonic pitch estimation.

- Utilizes a Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architecture to process spectral representations of audio for predicting note events and pitch activations.
- Training data includes various public datasets such as MAESTRO, Slakh, GuitarSet, Phenicx, iKala, and MedleyDB.

1.2.2 High-Resolution Piano Transcription and Pedal Regression Model [4]

- Focuses on high-resolution piano transcription, including the transcription of pedal events.
- Employs continuous regression methods to estimate precise note start and end times, enhancing transcription accuracy.
- Trained using the MAESTRO v2.0.0 dataset.

Key Features

- **Spectral Processing**: Converts audio signals into spectrograms to capture frequency and temporal information.
- **CNN-RNN Architecture**: CNN layers extract spatial features from spectrograms, while RNN layers model temporal dependencies.
- **Dual Output Heads**: One head predicts note events (start, end, pitch, velocity), and the other predicts sustain pedal activations.
- **Data Augmentation**: Techniques such as pitch shifting, time stretching, and adding background noise are applied to improve model generalization.

Chapter 2

Datasets

2.1 Evaluation Data

This system is evaluated on the following datasets:

1. **MAESTRO v3.0.0 Test Set** [**3**]: 177 audio files.
2. **MAPS Dataset** [**2**]: 60 audio files from the ENSTDkCl/MUS and ENSTDkAm/MUS subsets.
3. **SMD-Piano Dataset Version 2** [**5**]: 50 audio files.

2.2 Data Usage

- **Evaluation Data**: The evaluation datasets (MAESTRO test set, MAPS, and SMD) are not used for model training or tuning.
- **Validation Set**: The MAESTRO validation set is used solely for model development.

Bibliography

[1] Rachel M. Bittner, Juan José Bosch, David Rubinstein, Gabriel Meseguer-Brocal, and Sebastian Ewert. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022.

[2] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing*, 18(4):1124–1136, 2010.

[3] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[4] Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, and Yuxuan Wang. High-resolution piano transcription with pedals by regressing onsets and offsets times. *arXiv preprint arXiv:2010.01815*, 2020.

[5] Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. Saarland music data (SMD). 2011.