Brandon Carone, Cognition and Perception ‘26

Kenny Huang, Music Technology ‘23

Richa Namballa, Music Technology ‘23

MPATE-GE 2623: Music Information Retrieval

Dr. Magdalena Fuentes

**MIR Final Project Dataset**

**Goal**

Create a new dataset of jazz piano MIDI features with annotated chords that can be used for chord progression prediction tasks. If time permits, implement a basic Hidden Markov Model to test the application of the dataset to a task.

**Dataset to Annotate**

Jazz Piano MIDI Multitracks

Size: 310 MIDI files

Source: <https://bushgrafts.com/midi/>

**Reasoning**

The proposed dataset will be derived from [Doug McKenzie’s](https://bushgrafts.com/midi/) website and will feature 310 jazz MIDI multitrack files with piano as the melody line.

**Dataset Overview**

* 310 MIDI files
* Since it is MIDI data and not audio files, we are not sure if the ‘hours’ question applies here since the tempo can be adjusted to whatever we please, but the files equate to approximately 20 hours of content.
* The MIDI files are multitracks that vary in instrumentation. A piano is present in each file, so we will begin with annotating using the piano only and then extend to multitrack if time permits.

**Dataset Description**

* In our project, we aim to find a method to automate the chord annotation process for each file, as opposed to manually labeling every chord in each song. We will implement this process in a Python program.
  + If we include timestamps, we will use the tempo to derive the length of each bar in seconds (or milliseconds).
  + We will create a NumPy array for each song which includes each column representing one beat with the notes that the piano plays within that beat to estimate which chord is being played.
  + Each array will have accompanying metadata associated with it, including the title, instrumentation, key, tempo, etc.
  + Given the obscurity surrounding some lesser-known jazz chords, the annotation procedure will be human-in-the-loop. Any chord the program deems to be ambiguous will be flagged and the user will be able to manually fix the chord label for that section.
* We hope to be able to augment the dataset by segmenting each file into the different chord progressions that are found within it, rather than only providing one progression per file.
* Another method to augment the dataset would be to use the *pychord* library to transpose chords into other keys so that there are multiple versions of the same song.

**Deliverables**

* A Python program that can automatically annotate the jazz chord progressions for a MIDI file. For any ambiguous chords or sections which the program cannot determine an annotation for, the program will flag them and the human user can manually enter the corresponding label to ensure the annotation accuracy.
* Using our program, we will deliver all of the annotations for the 310 MIDI files with each annotation containing an array of each progression which is time aligned by beat and the associated metadata (title, key, tempo, etc.). The MIDI files will be linked to where we originally obtained them.
* The tool will be properly documented and available for use in a GitHub repository, labeled for educational purposes only.