LARGE-SCALE CONTENT-BASED MATCHING OF MIDI AND AUDIO FILES

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

MIDI files, when paired with corresponding audio recordings, can be used as ground truth for many music information retrieval tasks. We present a system which can efficiently match and align MIDI files to entries in a large corpus of audio content without using any metadata. The core of our approach is a neural network-based cross-modality hashing scheme which transforms acoustic feature matrices and MIDI piano rolls into sequences of vectors in a common Hamming space. Once represented in this way, we can efficiently perform large-scale dynamic time warping searches to match MIDI data to audio recordings. We evaluate our approach on the task of matching a huge corpus of MIDI files to the Million Song Dataset.

1. TRAINING DATA FOR MIR

Central to the task of content-based Music Information Retrieval is the curation of ground-truth data for tasks of interest (e.g. timestamped chord labels for automatic chord estimation, beat positions for beat tracking, prominent melody time series for melody extraction, etc.). The quantity and quality of this ground-truth is often instrumental in the success of MIR systems which utilize it as training data. Unfortunately, creating appropriate labels for a recording of a given song by hand often requires person-hours on the order of the length of the song. This often arguably makes the available training data a bottleneck to success for a given content-based MIR task.

It has previously been observed that MIDI files, when time-aligned to corresponding audio recordings, can be used to infer ground-truth information about a given song [3, 7]. This is due to the fact that a MIDI files can be viewed simplistically as a timed sequence of note annotations or a piano roll. It is much more straightforward to estimate, e.g., beat locations, chord labels, and the predominant melody from these representations than one which was derived from an audio signal. Unsurprisingly, a handful of tools have been developed for inferring this information from MIDI files [1, 2, 5, 6].

In [4], it is argued that some of the biggest successes in machine learning are thanks to the fact that "...a large training set of the input-output behavior that we seek to automate is available to us in the wild." The main motivation behind this project is that this crucial availability of data holds true for MIDI files - through a large-scale web scrape, we obtained 250,000 unique MIDI files, which is orders of magnitude larger than the datasets typically used for MIR research. We believe this proliferation of data is largely caused to two factors: First, that karaoke files are typically distributed as MIDI data and that karaoke is wildly popular, and second, that transcribing popular music as MIDI files was a common pastime for hobbyist musicians in the nineties.

1.1 Wrangling MIDI Files

The mere existence of a large collection of MIDI data is not enough, however. As noted above, in order to use MIDI files as ground truth, they need to be both matched (paired with a corresponding audio recording of the same song) and aligned (adjusted so that the timing of the events transcribed in the file match the audio recording). The latter problem has seen a great deal of research effort [3, 7], and will not be a main focus of this work.

Given a large corpora of audio and MIDI files, the task of matching entries from each may seem to be a trivial problem involving fuzzy text matching of the files' metadata. However, MIDI files have no formal mechanism for storing metadata (apart from text meta events, which are rarely used), and as a result the best-case scenario is that the artist and song title are included in the filename or subdirectory. While we found some examples of this in our collection of scraped MIDI files, the vast majority of the files had effectively no metadata information. Figure 1 shows a random sampling of subdirectory and filenames from our collection.

2. REFERENCES

[1] Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Conference on Music Information Retrieval*, pages 637–642, 2010.

[2] Tuomas Eerola and Petri Toiviainen. MIR in Matlab: The MIDI toolbox. In *Proceedings of the 5th Inter-*

```
J/Jerseygi.mid
V/VARIA18O.MID
Carpenters/WeveOnly.mid
2009 MIDI/handy_man1-D105.mid
G/Garotos Modernos - Bailanta De Fronteira.mid
Various Artists/REWINDNAS.MID
GoldenEarring/Twilight_Zone.mid
Sure.Polyphone.Midi/Poly 2268.mid
d/danza3.mid
100%sure.polyphone.midi/Fresh.mid
rogers_kenny/medley.mid
2009 MIDI/looking_out_my_backdoor3-Bb192.mid
```

Figure 1. Random sampling of 12 MIDI filenames and their parent directories from our corpus of 250,000 MIDI files scraped from the Internet.

national Conference on Music Information Retrieval, pages 22–27, 2004.

[3] Sebastian Ewert, Meinard Muller, Verena Konz, D Mullensiefen, and Geraint A Wiggins. Towards cross-version harmonic analysis of music. *Multimedia, IEEE Transactions on*, 14(3):770–782, 2012.

[4] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[5] Cory McKay and Ichiro Fujinaga. jSymbolic: A feature extractor for MIDI files. In *Proceedings of the International Computer Music Conference*, pages 302–305, 2006.

[6] Colin Raffel and Daniel P. W. Ellis. Intuitive analysis, creation and manipulation of MIDI data with `pretty_midi`. In *Proceedings of the 15th International Conference on Music Information Retrieval Late Breaking and Demo Papers*, 2014.

[7] Robert J Turetsky and Daniel PW Ellis. Ground-truth transcriptions of real music from force-aligned midi syntheses. *ISMIR 2003*, pages 135–141, 2003.