A DATA-CLEANSING FRAMEWORK FOR AGGREGATING ANNOTATED DATASETS FROM MIREX AUTOMATED CHORD ESTIMATION ARCHIVES

Jeff Miller **Johan Pauwels** **Mark Sandler**
Centre for Digital Music, Queen Mary University of London
{j.k.miller, j.pauwels, mark.sandler}@qmul.ac.uk

ABSTRACT

Identification and availability of suitable data sources is a well-known difficulty in music information retrieval research. For studies requiring annotated data, this can be compounded by inconsistent presentation formats, differences in methodologies, and annotation errors. By building a framework to apply automated data cleansing and standardization techniques to a collection of MIREX evaluation output data, we were able to extract a large, labelled chord data set for use in a harmonic modelling study.

1. INTRODUCTION

As part of a study on harmonic function modelling, it was necessary to examine the effects of subjectivity and discrepancies in chord label data. The study required empirical evidence, to be gathered by comparing chord transcriptions of audio recordings. To avoid bias and ensure a harmonically comprehensive model, chord label data from a number of musical pieces possessing a suitable variety of keys, chord vocabularies, and harmonic structures was required. Furthermore, for each recording multiple chord transcriptions were needed. These could have been produced by either humans or algorithms.

The Music Information Retrieval Evaluation eXchange (MIREX) Automated Chord Estimation (ACE) task has been running since 2008 and has been active in its present form since 2013[1-2]. Each year, submitted ACE algorithms are run on a standardized framework and are evaluated using various performance metrics. The algorithms are run on a number of musical audio datasets which remain the same (with some additions) from year to year. The result as of 2020 is a collection of over 35,000 files containing chord transcriptions for over 1,000 distinct audio recordings. This constitutes a substantial potential resource. The constructive reuse of this data has been explored by Ni et al. [3] as well as Koops et al. [4].

For the purposes of our study, several challenges had to be overcome before the MIREX ACE data could be used. These included both standard data cleansing considerations as well as domain-specific text processing of musical chord descriptions. The framework was implemented in Python 3.7. The code and an example output dataset are available at https://github.com/jeffkmiller/chord-data-aggregator.

2. PROCESS

Due to the large number of files to be scanned and harvested, it was essential that the process of data ingestion and formatting would be automated and run with minimal human intervention. This was also necessary to reduce the risk of process error. The data loader would need to filter out any files that did not contain chord label data. Some ACE algorithms have been tested repeatedly without modifications, thereby creating multiple files containing identical chord transcriptions. As this would result in over-representation of that data, these duplicates needed to be identified and removed. This was complicated by the fact that duplicated transcriptions may have different filenames, whereas two files with the same name could actually contain different transcriptions. Robust duplicate removal was accomplished through a series of cascading hash comparisons in a process which was substantially faster than a brute force search and compare algorithm.

The parent study required side-to-side framewise comparison of chord estimations from multiple transcriptions of each song. Although the MIREX output format is time-aligned, only the beginning and ending values are recorded for each chord. It was therefore necessary to expand these values to granular time frames so that the chord estimate for each standard time unit could be represented on a grid. A time frame step value of 100 milliseconds was chosen as a compromise which would allow a high degree of musically meaningful precision while keeping the datasets to a manageable size.

Although the MIREX submission guidelines specify the output chord labels should be formatted using the syntax proposed by Harte et al.[5] the resulting data collection still exhibits significant variations in format and scope which must be dealt with before the data can be integrated. Each of the 12 pitches in the equal temperament system can be described using multiple letter names and modifiers (such as "sharp" and "flat", also referred to as "accidentals"). For direct comparison of labels, the enharmonic spellings of the chord roots must be standardized.

A reference dictionary of preferred enharmonic spellings was created; the rules applied were: 1) accidentals in the root name should be avoided if possible, and 2) the key signature associated with the resulting chord should have the minimum number of accidentals. Thus, the root "B" would be preferable to "Cb", and "Db major" (5 flats) would be preferable to "C# major" (7 sharps). This corresponds generally to accepted Western musical practice.

Once the chord roots had been corrected, a similar process was required to standardize the chord types and their presentation formats. Harte's syntax allows for both longhand and shorthand presentations. In practice, the formats displayed in the MIREX results use various combinations of these presentations that must be homogenized to produce distinct descriptions for each chord type. In addition, there is considerable variety in the complexity of the chord vocabularies considered by the various algorithms. Some algorithms restrict their possible estimations to a small subset of simple chords to reduce errors. Others use larger vocabularies of chord types, which can result in a more detailed transcription at the cost of introducing a higher percentage of errors. As the purpose of the study was to investigate the nature of these discrepancies, it was necessary to map the various chord descriptions to the closest appropriate chord types, while retaining as much harmonic detail as possible. Information describing chord inversions and voicings was discarded, while the information regarding chord root and type was retained. A small number of transcriptions were presented in longhand format using only the chord root and parenthetical tuples of detected chord elements. As mapping of purely parenthetical chord descriptions was beyond the scope of this project, a data-cleansing process was applied to detect and discard these files.

Both the format cleansing and vocabulary matching were accomplished by parsing the labels for type information and comparing the parsed chord type descriptors to a chord vocabulary. Matched chord types were re-written using a standard format; un-matched chords were marked as such to allow statistical analysis at a later stage.

Finally, all the cleansed and standardized chord label records were automatically aggregated and written to a single file for ease of further analysis. The comma-separated file format was chosen as it is platform-agnostic, allowing the files to be imported into a wide variety of software for further processing and analysis. For purposes of data auditing and to aid the investigation of anomalies, an index linking each chord sequence to its generating algorithm and original source recording was automatically written to a separate file.

3. RESULTS

Using the described framework, we were able to repurpose existing data from the MIREX ACE task archives to create a data set of aggregated chord labels. The standardization of both presentation format and musical descriptions of the annotations made it possible to perform meaningful statistical analysis of the chord vocabularies, their distributions, and discrepancies between various algorithms as they attempted to describe each chord. The automation of the process made it possible to quickly modify parameters and reproduce data sets while maintaining data consistency and reducing process errors.

4. FUTURE WORK

Each exported .csv file contains standardized and time aligned chord data for one recording. At this time, compilation of chord data from multiple recordings requires further processing. Future work will include expanded chord vocabularies, the ability to map longhand parenthetical chord descriptions, and the development of an automated framework to aggregate chord estimation data from an entire collection of recordings into a single data file. This augmentation will result in a more comprehensive chord model and will also make it possible to directly compare a wider range of musical pieces and recordings.

5. REFERENCES

[1] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoust. Sci. Technol.*, vol. 29, no. 4, pp. 247–255, 2008.

[2] J. Pauwels and G. Peeters, "Evaluating automatically estimated chord sequences," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 749–753, 2013.

[3] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie, "Understanding effects of subjectivity in measuring chord estimation accuracy," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 12, pp. 2607–2615, 2013.

[4] H. V. Koops, W. Bas de Haas, D. Bountouridis, and A. Volk, "Integration and quality assessment of heterogeneous chord sequences using data fusion," *Proc. 17th Int. Soc. Music Inf. Retr. Conf. ISMIR 2016*, pp. 178–184, 2016.

[5] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolic Representation Of Musical Chords: A Proposed Syntax For Text Annotations." pp. 66–71, 2005.