Lead Sheet Transcription

Pierre Lardet

Abstract

This skeleton demonstrates how to use the `infthesis` style for undergraduate dissertations in the School of Informatics. It also emphasises the page limit, and that you must not deviate from the required style. The file `skeleton.tex` generates this document and should be used as a starting point for your thesis. Replace this abstract text with a concise summary of your report.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Pierre Lardet*)

Acknowledgements

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Lead sheets are a common way of representing music in a simple and concise format. They consist of an aligned melody and harmony through chord symbols. Lead sheets are widely used in the music industry, as they provide a quick and easy way to learn and perform music, espeically in jazz music where improvisation is common. Lead sheets are also used in music education, as they provide a simple way to introduce students to music theory and notation. However, creating lead sheets can be a time-consuming and error-prone process, especially for complex pieces of music. This project aims to develop a model that can automatically generate lead sheets from audio recordings of music, which could be used to help musicians, educators, and students create lead sheets more easily and accurately.

However, there is no good source of lead sheets for many songs. Popular websites such as Ultimate Guitar [XX] and HookTheory [XX] provide user-submitted chord annotations for almost any song, but the quality of these annotations can vary significantly [X]. Tools like Chordify [XX] can automatically generate chord annotations from audio recordings, but lack aligned melodies and have formats designed for complete beginners. Online forums where a variety of leads can be found do exist, such as MusicNotes [XX] and SheetMusicDirect [XX], but are expensive, have limited selections and are also user submitted, meaning their quality cannot be guaranteed.

To this end, we investigate the use of machine learning models for lead sheet generation, requiring the model to perform automatic chord recognition and melody transcription, and align them in order to generate a lead sheet.

1.2 Aims

The aims of this project are:

- To develop a model that can generate lead sheets from audio recordings of music.

- To implement and evaluate current state-of-the-art models for automatic chord recognition and melody transcription.

- Investigate the use of different data representations for music transcription and generation.

- To evaluate the performance of the model on a variety of music genres and styles.

- To investigate the use of different models/datasets/data augmentation methods for automatic chord recognition, melody transcription and lyric transcription.

- To build a useable, reproducible tool that can generate lead sheets from audio recordings of music.

1.3 Outline

The report is structured as follows:

- **Chapter 2** provides background information on lead sheets, music transcription, and related work.

- **Chapter 3** describes the experiments conducted in this project, including the data, models, training, and evaluation.

- **Chapter 4** presents and analyses the results of the experiments.

- **Chapter 5** concludes the report and provides suggestions for future work.

Chapter 2

Background & Related Work

2.1 Background

2.1.1 Lead Sheets

Lead sheets are a form of musical notation that contain both the melody and harmony. Only the dominant melody is present which is often the vocal line and is written in standard musical notation on a stave. The harmony is represented by chord symbols, which are written above the stave at the relevant time. A lead sheet also contains the key, time signature and, optionally, the lyrics of a song. An example of a lead sheet is shown in Figure 2.1.

Figure 2.1: An example of a lead sheet for 'Yesterday' by the Beatles.

Lead sheets are most notably used in jazz music where their origins lie. They date back to the mid 20th century, and were originally called 'fake sheets' because they were used by musicians to 'fake' their way through a song [17]. Any jazz musician worth their salt owns a 'real book', so called to distinguish it from the fake books that were used in

the past. Real books contain lead sheets for hundreds of jazz standards, and are still an essential tool for any seasoned jazz musician.

More recently they have served as a useful tool for musicians in other genres too, such as pop and rock, who want to learn and perform songs quickly and easily. They allow efficient communication of the important elements of a song, without the need for a full score, encouraging further improvisation and personalisation of the song.

Lead sheets do not contain all the information of a full score, such as dynamics, articulation, or specific voicings of chords. Furthermore, they are not suitable for all types of music, such as classical music, where a full score is necessary to convey the composer's intentions, or rap music where the lyrics and beat are normally the most important elements. However, they are a useful tool for many musicians, and are a common way of representing music for both learning and performing.

2.1.2 Automatic Music Transcription

Automatic Music Transcription (AMT) is a field within Music Information Retrieva (MIR) that aims to construct models that convert musical audio into symbolic representations. [2] provide a comprehensive review of different forms of transcription. For lead sheet transcription, we are interested in their highest level of transcription: notation-level. We want to be able to take an audio recording of a song and generate a lead sheet that contains the melody and harmony of the song. This is a challenging task, as it requires the model to isolate and transcribe a melody, transcribe the chordal information, and then combine these two elements into a coherent lead sheet where the melody and harmony are aligned in time/beat.

There are three sub-fields of musical transcription that are of interest to lead sheet generation: Automatic Chord Recognition (ACR), Melody Transcription (previously called F0 estimation) and Beat Detection.

2.1.3 Music Features

Recorded music can be represented in a variety of ways as input to a machine learning model. The simplest way is to leave the data as a waveform: a time-series vector of amplitudes. Data in the raw audio domain has seen successful use in generative models such as Jukebox [11] and RAVE [5].

A very commonly used representation of general audio data, and specifically musical data, is the spectrogram. A spectrogram is a conversion of the time-series data into the time-frequency domain, calculated by a SFTF (short-time Fourier Transform) Spectrograms are commonly used in many audio processing tasks, such as speech recognition, music recognition [31] and music transcription, specifically polyphonic transcription [29]. However, as noted by Pauwels et al. [25], only logarithmic spectrograms have been used in ACR tasks, and linear spectrograms have been used in melody transcription tasks.

A common version of the spectrogram used in music transcription is the Constant-Q Transform (CQT), originally proposed by Brown [3]. The CQT is a version of a

spectrogram with frequency bins that are logarithmically spaced and bin widths that are proportional to the frequency. This is motivated by the logarithmic nature of how humans perceive pitch: a sine wave that is perceived as one octave higher than another has double the frequency. The CQT is used in many music transcription tasks, such as automatic chord recognition [15], and a popular alternative, Melspectrograms, have found use in melody transcription [29].

Chroma vectors are a 12-dimensional time-series representation, where each dimension corresponds to a pitch class. Each element represents the presence of each pitch class in the Western chromatic scale in a given time frame. Such features have been generated by deep learning methods [24] or by hand-crafted methods [22].

More recently, features extracted generative models have been used as input. The proposed benefit is that the vast quantities of data used to train these models allows for rich representations of the music. Chris Donahue and Liang [9] use features from JukeBox [11] to train a transformer [30] for both melody transcription and chord recognition.

2.2 Related Work

2.2.1 Lead Sheet Transcription

Older work has attempted to automatically produce lead sheets [27, 32]. They use hand-crafted feature extractions and probabilistic models for melody and chord classification. AMT has developed dramatically since this work which leaves much room for improvement.

[9] is the only recent work we found that produces a full lead sheet generation model. They propose the use of features extracted from middle layers if Jukebox [11] which were found to lead to the best performance in downstream tasks [7]. These features are used to train a transformer, with a primary focus on melody transcription. The model and code is available[1]. They use the same methodology to train an ACR model, and combine these with beat detection and engraving software to produce a lead sheet.

The authors found the use of these generative pre-trained features to be beneficial for melody transcription. However, they use community-submitted annotations from an online forum, presumably for lack of a better alternative, whose quality and variety may be lacking. The nature of this data means that the model is only trained on 24 second audio clips, which limits performance on longer segments of audio. The authors claim that the overall model performs well, especially in the verses and choruses of pop music where vocal lines are loud and clear, and that performance across genres is surprisingly good given the pop-centric dataset. However, further analysis is ommitted and if the dataset contains music with few songs with more complex harmonic structures such as jazz music, then the model will produce overly simplistic representations of the chords. Crucially, the work does not explore ACR models beyond simply copying their method and dataset used in melody transcription and there is no quantitative evaluation of

---

[1]https://github.com/chrisdonahue/sheetsage/tree/main

performance in ACR. Furthermore, the system can struggle with melody lines shifting between instruments, quiet vocal lines

2.2.2 Automatic Chord Recognition

[25] provides an overview of ACR since its inception in 1999 with the work of Fujishima [13] up to 2019, and provide suggestions for future avenues of research. Among them are the lack of exploration of feature and chord representations and the imbalance with chords classes present in chord datasets.

Datasets that have seen common use in ACR over the last decade include:

- Mcgill *Billboard*: 890 chord annotations of songs randomly selected from the Billboard 'Hot 100' Chart bnetween 1958 anmd 1991. [4]

- *Isophonics*: 300 annotations of albums by The Beatles, Carole King and Zweieck. [6]

- *RWC-Pop*: 100 pop songs with annotations available for chords [2]. [14]

- *USPop*: 195 annotations of a larger dataset with artists chosen for popularity. [1]

- *JAAH*: 113 annotations of a collection of jazz recordings. [12]

Other, small datasets, also exist, and have been compiled together into the *Chord Corpus* by de Berardinis et al. [10], with standardised annotation formats.

A variety of machine learning methods have been applied to such datasets. Older methods such as [20] use HMMs, but more recent methods use deep learning. Often, CNNs and RNNs are used in combination [33, 18, 23] with CNNs performing feature extraction from a spectrogram feature, and an RNN (normally Bi-LSTM or Bi-GRU) predicts chord sequences. More recently, transformers have been applied to the entire process Chris Donahue and Liang [9] and Chen and Su [8].

Evaluation is typically done using accuracy and recall of correct chord predictions, or more music-aware measures such as the correct root note, 3rd or the MIREX metric which measures that chords have at least 3 notes in common. These are all implemented by Raffel et al. [26] in the `mir_eval` library [3]. Qualitative evaluation is also often carried out.

A big problem frequently encountered in ACR is the lack of labelled data. This work uses 1200 labelled songs with audio. This is due to the difficulty and time associated with labelling data aligned in time and the legal sensitivity of the data involved.Data has been scaled up using augmentation and semi-supervised learning [16] with some success. Research has been done into the use of synthetic data [19, 28] and supervised learning [21] for MIR tasks, but not for ACR.

Another problem is that the existing data is often imbalanced, with a large number of scommon chords like major and minor chords and fewer chords like diminished and augmented chords, or chords with upper extensions and inversions. This can lead to

---

[2]https://github.com/tmc323/Chord-Annotations
[3]https://mir-evaluation.github.io/mir_eval/

models that are biased towards predicting major and minor chords. Attempts to address this have been made by re-weighting classes [18] or adjusting the sampling of training examples to balance chord classes [24].

2.2.3 Melody/Vocal Transcription

Chapter 3

Experiments

3.1 Data

3.2 Models

3.3 Training

3.4 Evaluation

Chapter 4

Results

4.1 Quantitative Results

4.2 Qualitative Results

4.3 Lead Sheet Generations

Chapter 5

Conclusions

Bibliography

[1] Adam Berenzweig, Beth Logan, Daniel P. Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

[2] Bhuwan Bhattarai and Joonwhoan Lee. A comprehensive review on music transcription. *Applied Sciences*, 13:11882, 10 2023. doi: 10.3390/app132111882.

[3] Judith Brown. Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, 89:425–, 01 1991. doi: 10.1121/1.400476.

[4] John Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground truth set for audio chord recognition and music analysis. pages 633–638, 01 2011.

[5] Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis, 2021. URL https://arxiv.org/abs/2111.05011.

[6] Chris Cannam, Craig Landone, and Mark Sandler. Omras2 metadata project. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 309–310, 2009.

[7] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval, 2021. URL https://arxiv.org/abs/2107.05677.

[8] Tsung-Ping Chen and Li Su. Harmony transformer: Incorporating chord segmentation into harmony recognition. In *International Society for Music Information Retrieval Conference*, 2019. URL https://api.semanticscholar.org/CorpusID:208334896.

[9] John Thickstun Chris Donahue and Percy Liang. Melody transcription via generative pre-training, 2022. URL https://arxiv.org/abs/2212.01884.

[10] Jacopo de Berardinis, Albert Meroño-Peñuela, Andrea Poltronieri, and Valentina Presutti. Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs. *Scientific Data*, 10, 09 2023. doi: 10.1038/s41597-023-02410-w.

[11] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020. URL https://arxiv.org/abs/2005.00341.

[12] Gabriel Durán and Patricio de la Cuadra. Transcribing Lead Sheet-Like Chord Progressions of Jazz Recordings. *Computer Music Journal*, 44(4):26–42, 12 2020. ISSN 0148-9267. doi: 10.1162/comj_a_00579. URL https://doi.org/10.1162/comj_a_00579.

[13] Takuya Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *International Conference on Mathematics and Computing*, 1999. URL https://api.semanticscholar.org/CorpusID:38716842.

[14] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical, and jazz music databases. 01 2002.

[15] Eric J. Humphrey and Juan P. Bello. Rethinking automatic chord recognition with convolutional neural networks. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 357–362, 2012. doi: 10.1109/ICMLA.2012.220.

[16] Yun-Ning Hung, Ju-Chiang Wang, Minz Won, and Duc Le. Scaling up music information retrieval training with semi-supervised learning, 2023. URL https://arxiv.org/abs/2310.01353.

[17] 99 The real book, 2021. URL https://podcasts.apple.com/us/podcast/the-real-book/id394775318?i=1000519252462.

[18] Junyan Jiang, K. Chen, Wei Li, and Gus G. Xia. Large-vocabulary chord transcription via chord structure decomposition. In *International Society for Music Information Retrieval Conference*, 2019. URL https://api.semanticscholar.org/CorpusID:208334209.

[19] Nadine Kroher, Helena Cuesta, and Aggelos Pikrakis. Can musicgen create training data for mir tasks?, 2023. URL https://arxiv.org/abs/2311.09094.

[20] Kyogu Lee and Malcolm Slaney. Automatic chord recognition from audio using a hmm with supervised learning. pages 133–137, 01 2006.

[21] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. Mert: Acoustic music understanding model with large-scale self-supervised training, 2024. URL https://arxiv.org/abs/2306.00107.

[22] Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult chords. pages 135–140, 01 2010.

[23] Brian McFee and Juan Pablo Bello. Structured training for large-vocabulary chord recognition. In *International Society for Music Information Retrieval Conference*, 2017. URL https://api.semanticscholar.org/CorpusID:3072806.

[24] Jeff Miller, Ken O'Hanlon, and Mark B. Sandler. Improving balance in automatic chord recognition with random forests. In *2022 30th European Signal Processing*

Conference (EUSIPCO), pages 244–248, 2022. doi: 10.23919/EUSIPCO55093. 2022.9909558.

[25] Johan Pauwels, Ken O'Hanlon, Emilia Gómez, and Mark B. Sandler. 20 years of automatic chord recognition from audio. In *International Society for Music Information Retrieval Conference*, 2019. URL `https://api.semanticscholar. org/CorpusID:208334309`.

[26] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. Mir_eval: A transparent implementation of common mir metrics. In *International Society for Music Information Retrieval Conference*, 2014. URL `https://api.semanticscholar.org/CorpusID:17163281`.

[27] Matti P. Ryynänen and Anssi P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008. doi: 10.1162/comj.2008.32.3.72.

[28] Gakusei Sato and Taketo Akama. Annotation-free automatic music transcription with scalable synthetic data and adversarial domain confusion, 2024. URL `https://arxiv.org/abs/2312.10402`.

[29] Keisuke Toyama, Taketo Akama, Yukara Ikemiya, Yuhta Takida, Wei-Hsiang Liao, and Yuki Mitsufuji. Automatic piano transcription with hierarchical frequency-time transformer, 2023. URL `https://arxiv.org/abs/2307.04305`.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`.

[31] Avery Wang. An industrial strength audio search algorithm. 01 2003.

[32] Jan Weil, Thomas Sikora, Jean-Louis Durrieu, and Gaël Richard. Automatic generation of lead sheets from polyphonic music signals. pages 603–608, 01 2009.

[33] Yiming Wu, Tristan Carsault, and Kazuyoshi Yoshii. Automatic chord estimation based on a frame-wise convolutional recurrent neural network with non-aligned annotations. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019. doi: 10.23919/EUSIPCO.2019.8902741.

Appendix A

First appendix

A.1 First section