Piano Transcription from Audio-Playalongs

Practical Work in AI SS2023
Project Report

Sebastian Sonderegger
K1246236
Bachelor Student in AI
Johannes Kepler University Linz, Austria
s.sonderegger@mailbox.org
Supervisor: Francesco Foscarin, ICP

1 Project Outline

The aim of this project was to find, test and evaluate state-of-the art Music Information Retrieval (MIR) tools to obtain a system capable of transcribing raw audio (piano) music to symbolic music (MIDI), for further use in training generative systems with focus on Jazz Music. Among the tools tested, the most promising were the DEMUCS-Source Separator[1], the Piano Transcription Inference Transcriptor [2], and the Madmom Downbeat-Tracker[3], tested on Aebersold jazz backing-tracks [4] and evaluated on the Filosax dataset [5], kindly provided by David Foster and Maryland University .

2 The Data

In this project two main datasets were used, the Aebersold jazz backing-tracks Collection (1500 files) and the Filosax Dataset (48 files), which is actually a subset of the Aebersold Collection, thus mainly the hand-crafted annotations, provided by the authors were used. All tools tested in 3.1 were tested on samples from the Aebersold Collection, the Beat-Detector in 4.1 was evaluated with the Filosax annotations.

2.1 Data Preparation

What makes Aebersold backing-tracks optimal for this kind of experiments is that they always adhere to a fixed channel-layout, with drums centered, bass only on the left channel and piano only on the right. Thus to make it easier for the models to separate the piano source, the right channel (piano and drums) was extracted and centered.

3 Experiments

3.1 Audio Source Separation

Source Separation in the audio domain is the process of separating an audio track into so-called stems, where in the optimal case each resulting stem corresponds to an instrument audible in the track.

3.1.1 Spleeter (Deezer)

Spleeter is an Audio Source Separation tool developed by the online music platform Deezer and made available open-source. Spleeter uses pre-trained U-net models, U-nets are encoder/decoded Convolutional Neural Networks (CNNs).[6] It was mostly trained on the Bean dataset (24k of Pop/Rock songs with separated instrument tracks). Audio files can be separated into 2 (vocals, other), 4(vocals, bass, drums, other) or 5 stems (vocals, bass, drums, piano, other).

3.1.2 Demucs Hybrid Transformer

The open-source Hybrid Transformer Demucs model by Facebook Research is also comprised of U-net CNN's, but additionally some of the inner layers are replaced by Transformer layers with self- as well as cross-attention among domains.[1] Demucs separates audio into 4 stems (drums, bass and vocals and other) and it was trained on the MUSDB18 dataset plus additional 800 songs (not further specified).[7]

3.1.3 Results

After applying the models to various samples of the Aebersold collection to extract the piano source and comparing the results, there was subjective, but convincing evidence that the Demucs model performs better on the task. Namely the output was much clearer in sound, while Spleeter's output sounded 'damped' which indicates that higher frequencies of the piano sound were lost to the drum source, which produces noise-like frequencies over the whole spectrum. Also using Spleeter's 5-stems approach, which offers a specific piano stem, did rather worsen the result because some frequencies from the piano were lost to the "other" stem.

3.2 Audio Transcription

Audio Transcription from raw audio is the process of extracting onsets, pitch and length of every single note audible in a piece, and convert it to symbolic notation (MIDI or MusicXML), which then can again be used to generate music simliar to the original audio with f.e. prerecorded (sampled) MIDI-soundfonts.

3.2.1 Piano Transcription Inference (PTI)

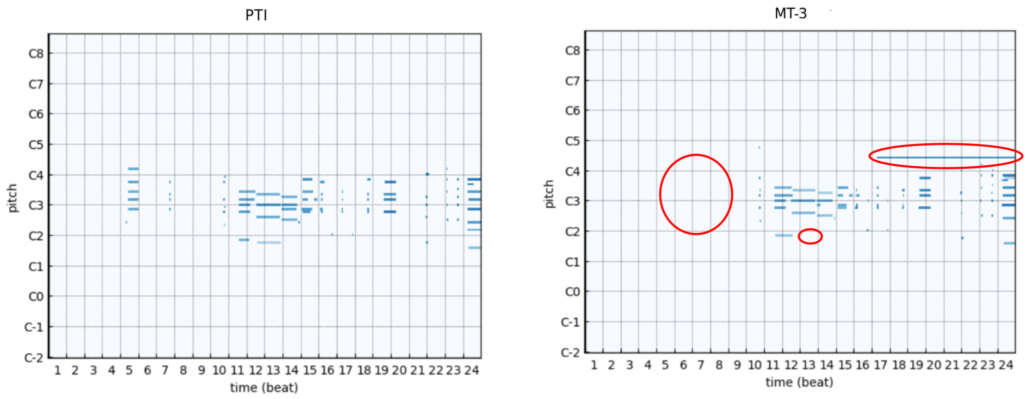
This regression-based model is specifically designed to transcribe piano music from audio to MIDI and is thus perfect for the task at hand.[8]

3.2.2 Magenta MT-3

MT-3 is a Transformer based multi-purpose transcription model by Magenta, which in a smaller version also includes a model solely for piano transcription which was used for the experiments (it was assumed that a model trained for recognizing various instruments would not perform better then one specifically trained on piano music).[9]

3.2.3 Results

Both models were applied to the Aebersold samples and then the resulting MIDI-file was again converted to audio with GeneralUser GS [10], a publicly available MIDI-soundfont, which was sampled form a real



(a) (b)

Figure 1: Comparison of MIDI-output from both models (excerpt of Filosax file No.44)

piano. The PTI model data worked surprisingly well from a first glance and comparing it to the output of MT-3 confirmed our sentiments, as the MIDI transcription produced by MT-3 contained a lot of "ghost-notes" not evident in the audio file and also whole parts were left out, while the PTI's output was mostly accurate. This is again a subjective conclusion, derived from cross-listening between the original audio and the transcription, as sufficient tools to compare the 'goodness' of two transcriptions do not (yet) exist and/or go beyond the scope of this project. It is possible though to compare the two MIDI outputs visually with each other, as you can do for yourself in Fig. 1.

4 Evaluation

4.1 Downbeat Detection - Madmom

Madmom is a pyhton-library containing different MIR-tools, one of which is a downbeat-tracker that also does offline downbeat detection. Madmom was developed by Sebastian Böck at JKU's Institute for Computational Perception (ICP). This downbeat-tracker is a multi-model system comprised of multiple Recurrent Neural Networks (RNNs) with a dynamic Bayesean network on top for model selection. [3] Evaluation was done with the Filosax data and the mir_eval library, which is the standard library for evaluating MIR tools.[11] Performance was evaluated using the standard, built-in, F-measure, which has a tolerance window of 7 frames, inside which beats are still considered correct. In the plot in Fig. 2 we see that performance was almost perfect (>0.9) for many test cases (Fig. 3). Common mistakes are offsets by one or more beats (Fig. 4) (unlike pop-music, jazz-music is often accentuated on the 2nd and 4th beat), or octave errors - wrong assumption of tempo, resulting in half or twice as many predicted beats (Fig. 5).

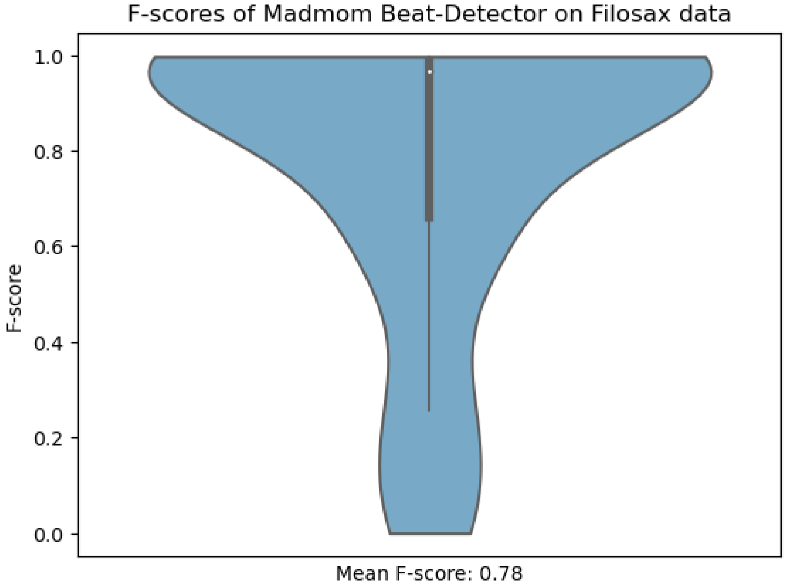


Figure 2: Results of the Madmom model on the Filosax data (total of 48 audio files)

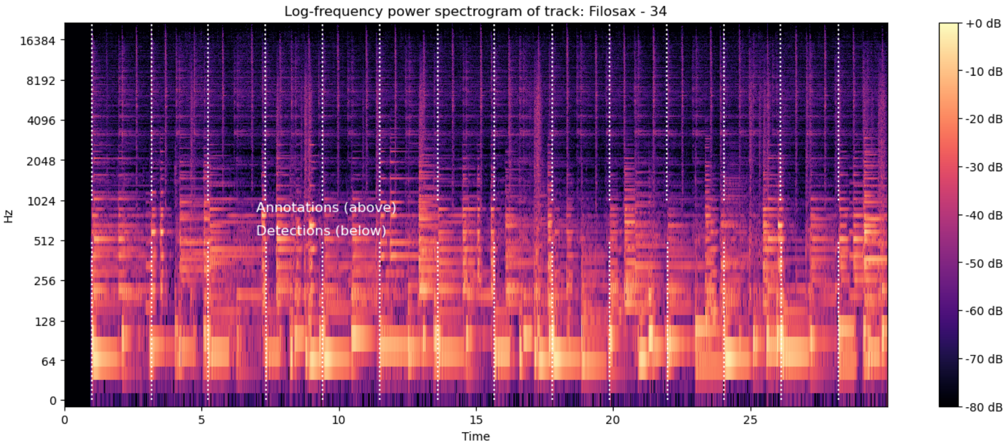


Figure 3: Example of a perfectly accurate prediction visualized on top of the spectrogram of the respective audio file (30 seconds of audio, downbeats are the white dotted lines).

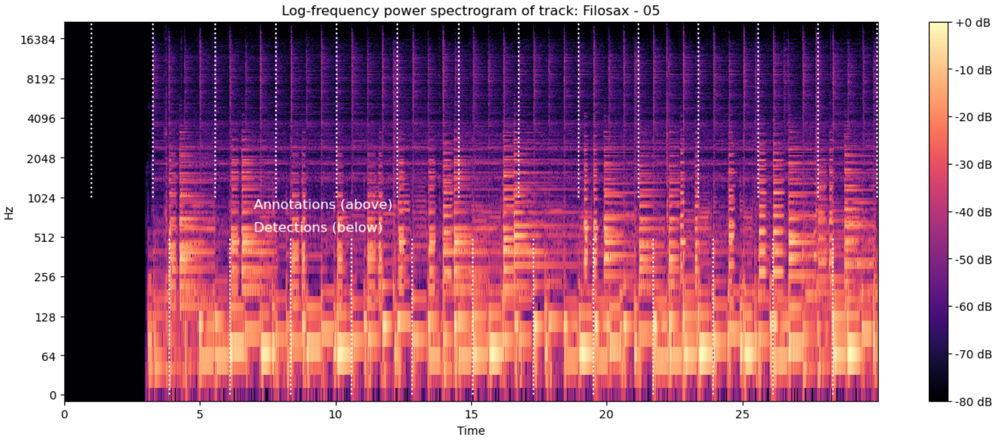


Figure 4: Example of a prediction offset by one beat.

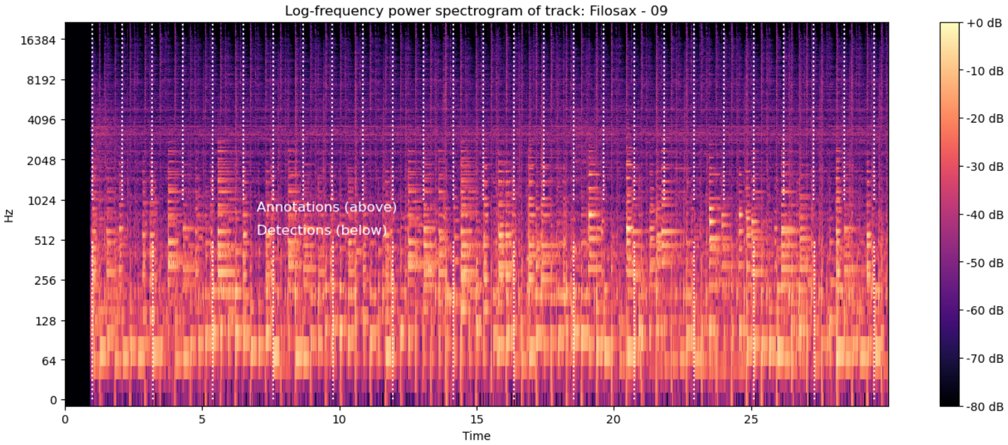


Figure 5: Example of an octave error (half of the real tempo was estimated)

5 Conclusions and Outlook

The process to get from raw audio to a symbolic representation involves many steps, each of which has it's own tools and some are more developed than others. Source separation, at least on 'simple' audio with only 2-3 different instruments, I would consider as an almost solved problem, Transcription on the other hand, I would not, as Multi-Pitch-Detection is still a hard problem, especially with multiple instruments present. Beat Detection tools like madmom are already very accurate, recent approaches also employ Transformers, which, like in many other sequence modeling problems, seems a promising way to go [12]. High ambiguity in music per se, does not make the task easier and in the end, not even humans can always agree on the same tempo or chords of a piece.

6 Code and Data

- The Aebersold Collection is not open-source and has to be purchased, see [4].
- Access to the Filosax Dataset can be requested online, at [13].
- To access the experiments conducted, as well as the notebook for Piano-to-MIDI transcription visit the github-repository [14]. Example file in the github repository is taken from [15]

References

Rouard, S., Massa, F., & Défossez, A. (2022). Hybrid transformers for music source separation.

Kong, Q., Li, B., Song, X., Wan, Y., & Wang, Y. (2021a). High-resolution piano transcription with pedals by regressing onset and offset times.

Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., & Widmer, G. (2016). Madmom: A new python audio and music signal processing library.

Jazz, J. A. (1967). *Aebersold*. Retrieved July 5, 2023, from https://www.jazzbooks.com/

Foster, D., & Dixon, S. (2021). Filosax: A dataset of annotated jazz saxophone recordings. In J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, & A. Srinivasamurthy (Eds.), *Proceedings of the 22nd international society for music information retrieval conference, ISMIR 2021, online, november 7-12, 2021* (pp. 205–212). https://archives.ismir.net/ismir2021/paper/000025.pdf

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: A fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, *5*, 2154. https://doi.org/10.21105/joss.02154

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019). MUSDB18-HQ - an uncompressed version of musdb18. https://doi.org/10.5281/zenodo.3338373

Kong, Q., Li, B., Song, X., Wan, Y., & Wang, Y. (2021b). High-resolution piano transcription with pedals by regressing onset and offset times.

Hawthorne, C., Simon, I., Swavely, R., Manilow, E., & Engel, J. (2021). Sequence-to-sequence piano transcription with transformers.

Collins, C. (n.d.). *General USer GS MIDI-Soundfont*. Retrieved July 5, 2023, from http://www.schristiancollins.com/generaluser.php

Raffel, C., Mcfee, B., Humphrey, E., Salamon, J., Nieto, O., Liang, D., & Ellis, D. (2014). Mir$_{e}val$: A transparent implementation of common mir metrics. *Proceedings - 15th International Society for Music Information Retrieval Conference (ISMIR 2014).*

Zhao, J., Xia, G., & Wang, Y. (2022). Beat transformer: Demixed beat and downbeat tracking with dilated self-attention.

Foster, M. (n.d.). *Filosax, access request*. Retrieved July 5, 2023, from https://zenodo.org/record/6335779

Sonderegger, S. (n.d.). *Piano Transcription - Project in AI*. Retrieved July 5, 2023, from https://github.com/sebson/project-ai

GmbH, P. (n.d.). *Pixabay, graphics and data you may use freely for editorial purpose*. Retrieved July 5, 2023, from https://pixabay.com/music/search/jazz