Automatic Music Transcription

Overview, Onsets and Frames, Unaligned Supervision

Matilde Tozzi

Ferienakademie 2024

September 2024

Table of contents

Overview

Definition

Automatic Music Transcription (AMT) is the design of computational algorithms to convert acoustic music signals into some form of music notation. [BenetosMusicTranscription]

Subtasks:

- multipitch estimation
- onset and offset detection
- instrument recognition
- beat and rhythm tracking
- dynamics
- score typesetting

Usual Workflow

- **(a)** audio waveform as input
- **(b)** time-frequency representation
- **(c)** piano-roll (MIDI: Musical Instrument Digital Interface) representation as output
- **(d)** typeset music score

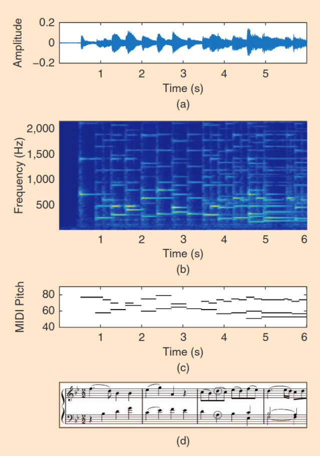

Figure 1: Source: [BenetosMusicTranscription] (Images courtesy of the MIDI Aligned Piano Sound database).

Key Challenges[1]

1. multiple simultaneous sources
2. harmonic relations in overlapping sounds
   - C major chord, fundamental frequency ratio C:E:G 4:5:6
   - harmonic overlap 46.7%, 33.3%, 60% for C, E, and G respectively
3. high synchronization of onsets and offsets between different voices $\Rightarrow$ no statistical independence between sources
4. annotation is very time consuming and requires high expertise
   - sheet music is not a good ground-truth: not time-aligned, not an accurate performance representation

[1] [BenetosMusicTranscription]

AMT Approaches

- **(a) frame level** = estimation of the number of and pitch of notes that are simultaneously present in each time frame ($\sim 10ms$), independently in each *time frame*
- **(b) note level** = connects pitch estimates over time into *notes* (pitch, onset time, offset time)
- **(c) stream level** (multipitch streaming) = grouping of estimated pitches or notes into *streams* (one instrument or musical voice)

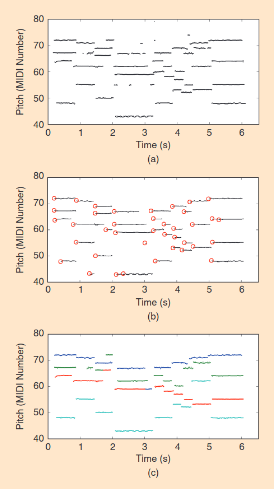

Figure 2: First phrase of J.S. Bach's chorale *Ach Gott und Herr*. Source: [BenetosMusicTranscription].

State of the Art I

1. **Negative Matrix Factorization** (not covered in this seminar) Represent a given nonnegative time-frequency representation $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{MxN}$ as a product of two nonnegative matrices: a **dictionary** $\boldsymbol{D} \in \mathbb{R}_{\geq 0}^{MxK}$ and an **activation matrix** $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{KxN}$. The goal is to minimize a distance (or divergence) between $\boldsymbol{V}$ and $\boldsymbol{DA}$ w.r.t. $\boldsymbol{D}$ and $\boldsymbol{A}$.

State of the Art II

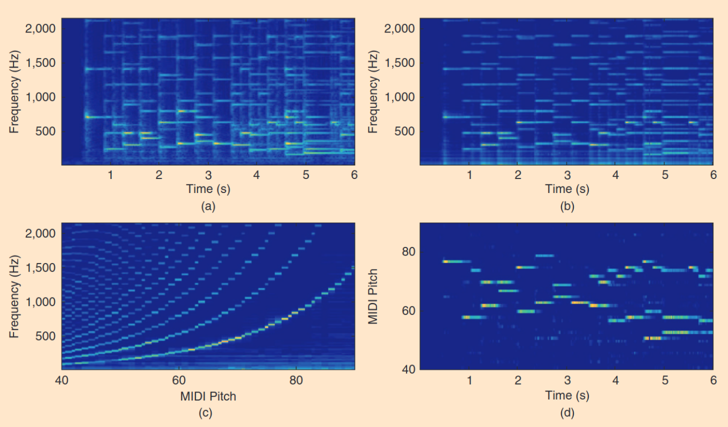

Figure 3: And example of NMF, using the same audio recording as in Figure 1: **(a)** input spectogram ***V***, **(b)** approximated spectogram ***DA***, **(c)** dictionary ***D***, and **(d)** activation matrix ***A***. Source: [BenetosMusicTranscription].

State of the Art III

2. **Neural Networks** (focus of this seminar)
   The most popular approach of this type is called **Onsets and Frames**, because it consists of two chained NNs. One detects note onset, and its outuput is used to inform a second network that focuses on perceiving the note lengths.

Bibliography and Acknowledgements I

A special thanks to **Abhirup Saha** for the slides on the topic.

[BenetosMusicTranscription] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, “Automatic Music Transcription: An Overview,” IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 20–30, Jan. 2019, doi: `https://doi.org/10.1109/msp.2018.2869928`.

[HawthorneOnsetsFrames] C. Hawthorne et al., “Onsets and Frames: Dual-Objective Piano Transcription,” International Symposium/Conference on Music Information Retrieval, pp. 50–57, Sep. 2018, doi: `https://doi.org/10.5281/zenodo.1492341`.

Bibliography and Acknowledgements II

[MamanUnalignedAMT] B. Maman and A. Bermano, “Unaligned Supervision for Automatic Music Transcription in-the-Wild.” Accessed: Sep. 19, 2024.