

Automatic Music Transcription

Overview, Onsets and Frames, Unaligned Supervision

Matilde Tozzi

Ferienakademie

September 2024

- 1 Overview
 - Definition
 - Usual Workflow
 - AMT Approaches
 - State of the Art
 - Key Challenges

- 2 Unaligned Supervision for AMT in the Wild
 - Scheme

Overview

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

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

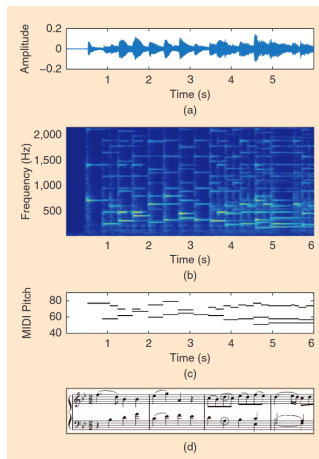


Figure 1: Source: [BenetosMusicTranscription] (Images courtesy of the MIDI Aligned Piano Sound database).

- **(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)

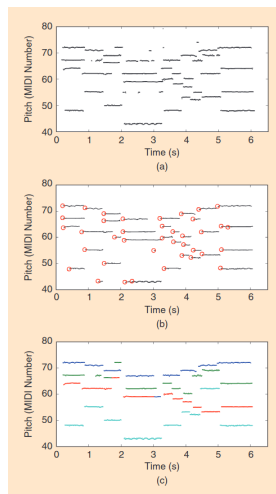


Figure 2: First phrase of J.S. Bach's chorale *Ach Gott und Herr*. Source: [BenetosMusicTranscription].

Neural Networks

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 output is used to inform a second network that focuses on perceiving the *note lengths*.

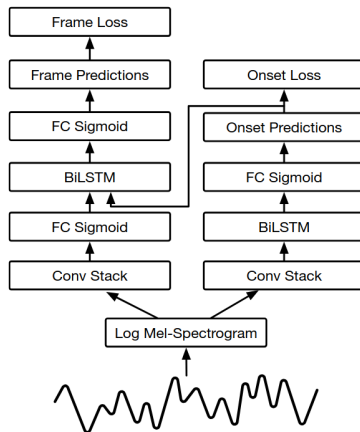


Figure 3: Source: [HawthorneOnsetsFrames].

Mel-Spectrogram

The **mel scale** (after the word *melody*) is a perceptual scale of pitches judged by listeners to be equal in distance from one another.

- Reference point: 1000 mels = 1000 Hz tone, 40 dB above the listener's threshold.
- Above about 500 Hz, increasingly large intervals are judged by listeners to produce equal pitch increments.

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$

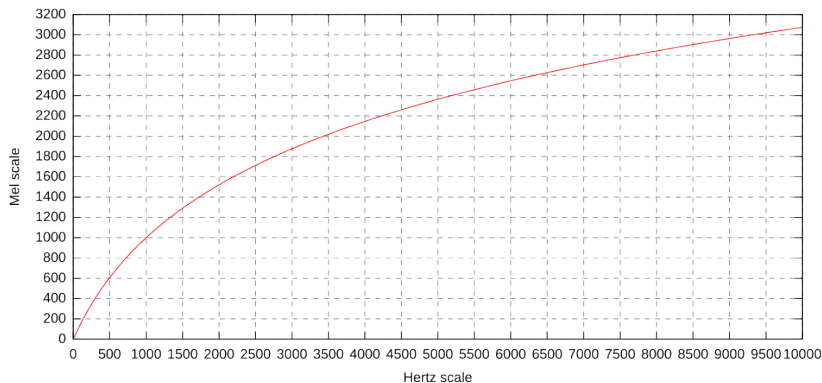


Figure 4: Relation between Hertz and Mel scales. Source: [MelScale].

- 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

Examples of metric limitations for Onsets and Frames¹

Original Score

Note timing jittered, but still within tolerance

Many 1-frame notes added

¹[HawthorneOnsetsFrames]

²[BenetosMusicTranscription]

Unaligned Supervision for AMT in the Wild

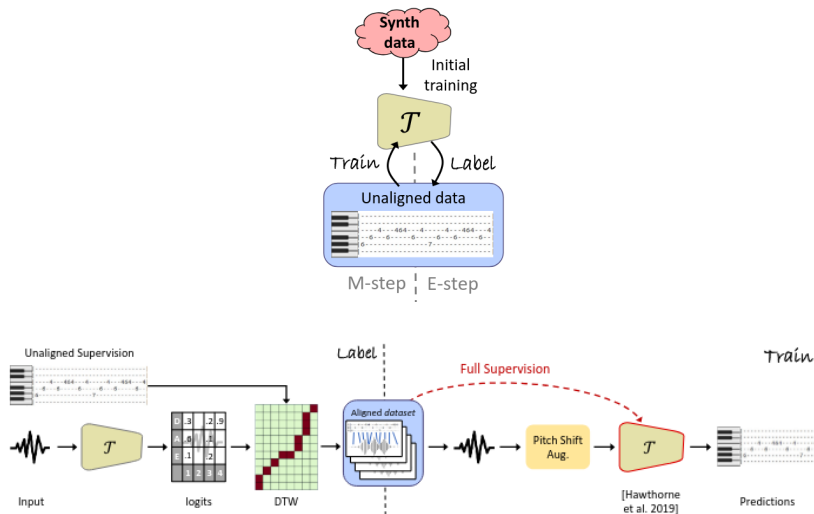


Figure 5: Source: [MamanUnalignedAMT].

- [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>.
- [MamanUnalignedAMT] B. Maman and A. Bermano, “Unaligned Supervision for Automatic Music Transcription in-the-Wild.”
- [MelScale] https://en.wikipedia.org/wiki/Mel_scale