Automatic Music Transcription

Overview, Onsets and Frames, Unaligned Supervision

Matilde Tozzi

Ferienakademie

September 2024

Definition

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

Subtasks:

- note *onset* and *offset* detection
- *instrument* recognition
- *beat* and *rhythm* tracking
- ...

Usual Workflow

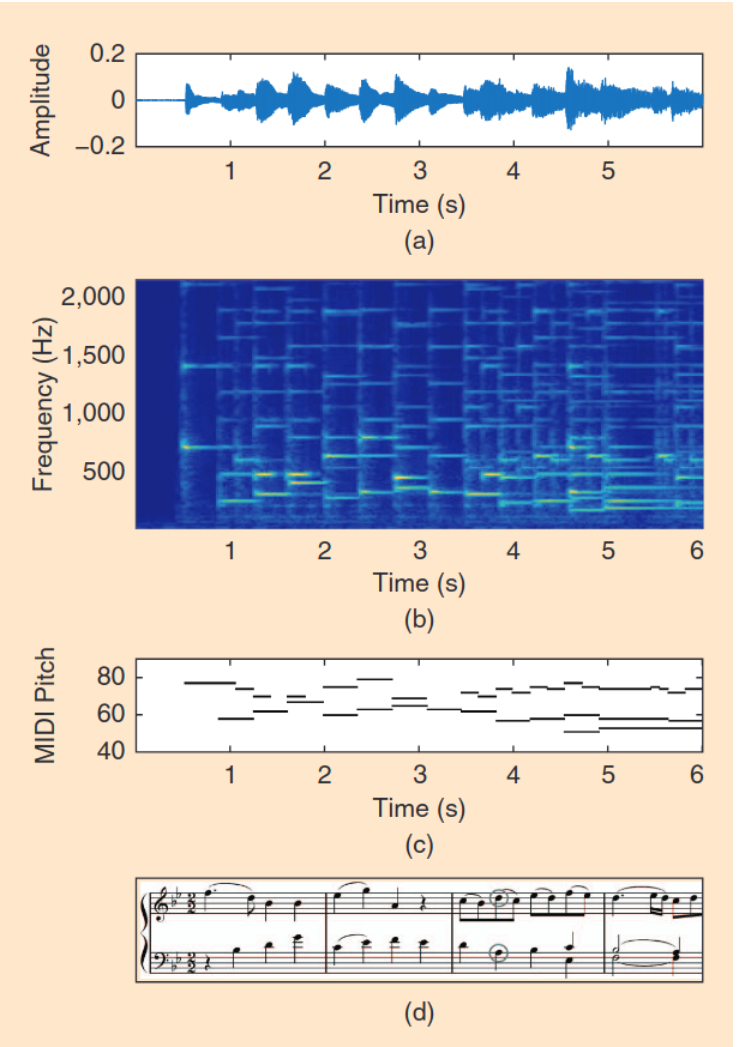

Figure 1: Source: [BenetosMusicTranscription] (Images courtesy of the MIDI Aligned Piano Sound database).

AMT Approaches

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

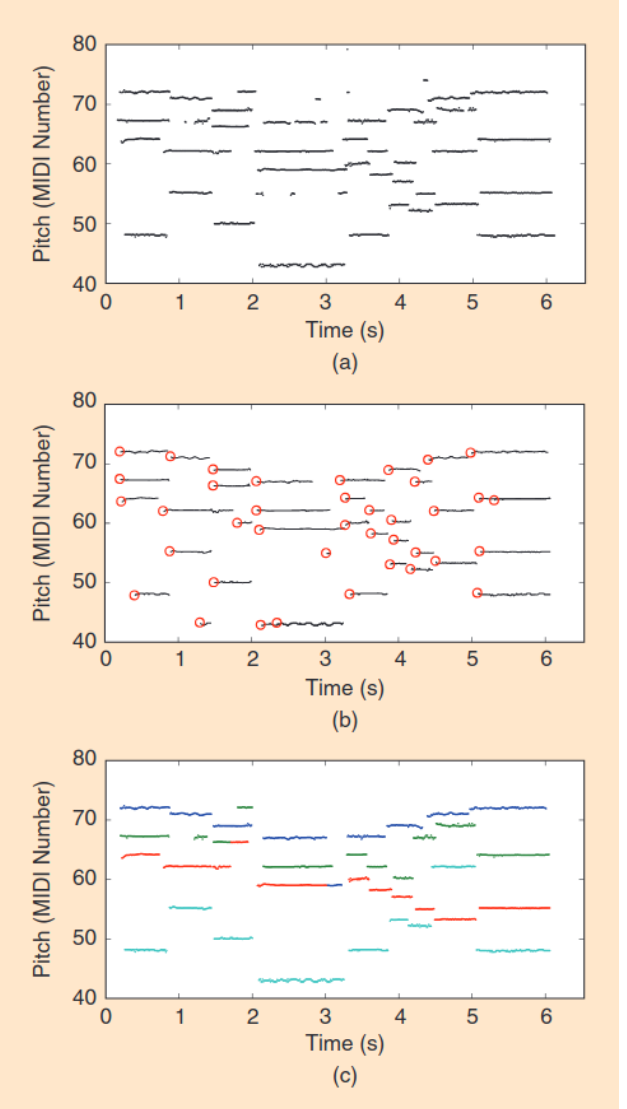

Figure 2: First phrase of J.S. Bach's chorale *Ach Gott und Herr*. Source: [BenetosMusicTranscription].

State of the Art I

Onsets and Frames

Two chained **Neural Networks**:

1. detect *note onset*
2. perceive *note lengths* (frames)

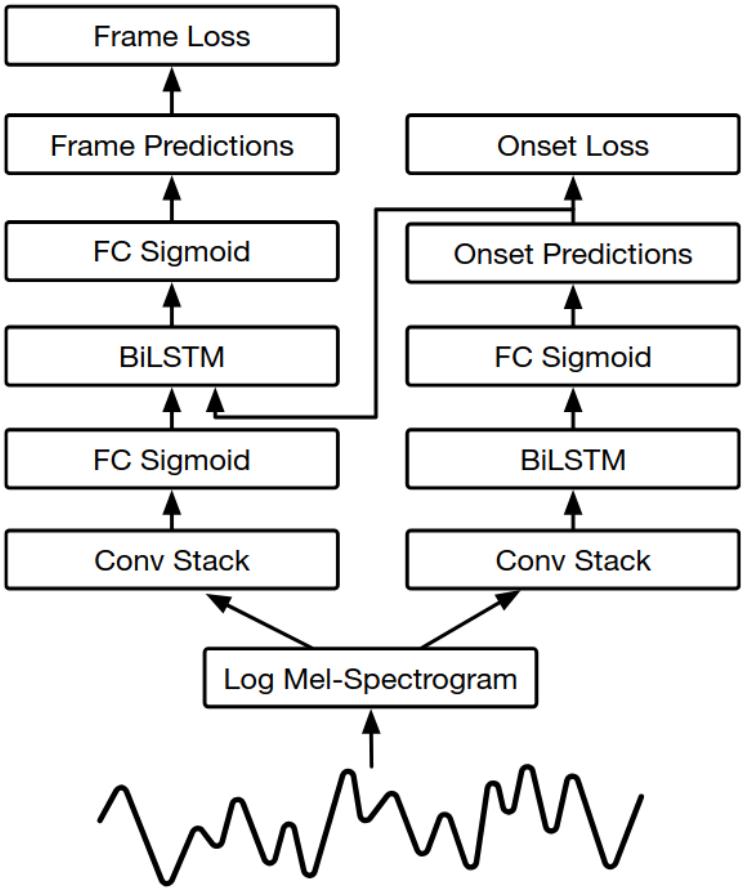

Figure 3: Source: [HawthorneOnsetsFrames].

State of the Art II

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

[MelScale]

Key Challenges[2]

1. **harmonic relations** in overlapping sounds
2. **high synchronization** of onsets and offsets between different voices $\Rightarrow$ no statistical independence between sources
3. **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 Frames[1]

Original Score

Note timing jittered, but still within tolerance (50ms)

Many 1-frame notes added

[1] [HawthorneOnsetsFrames]
[2] [BenetosMusicTranscription]

Unaligned Supervision for AMT in the Wild I

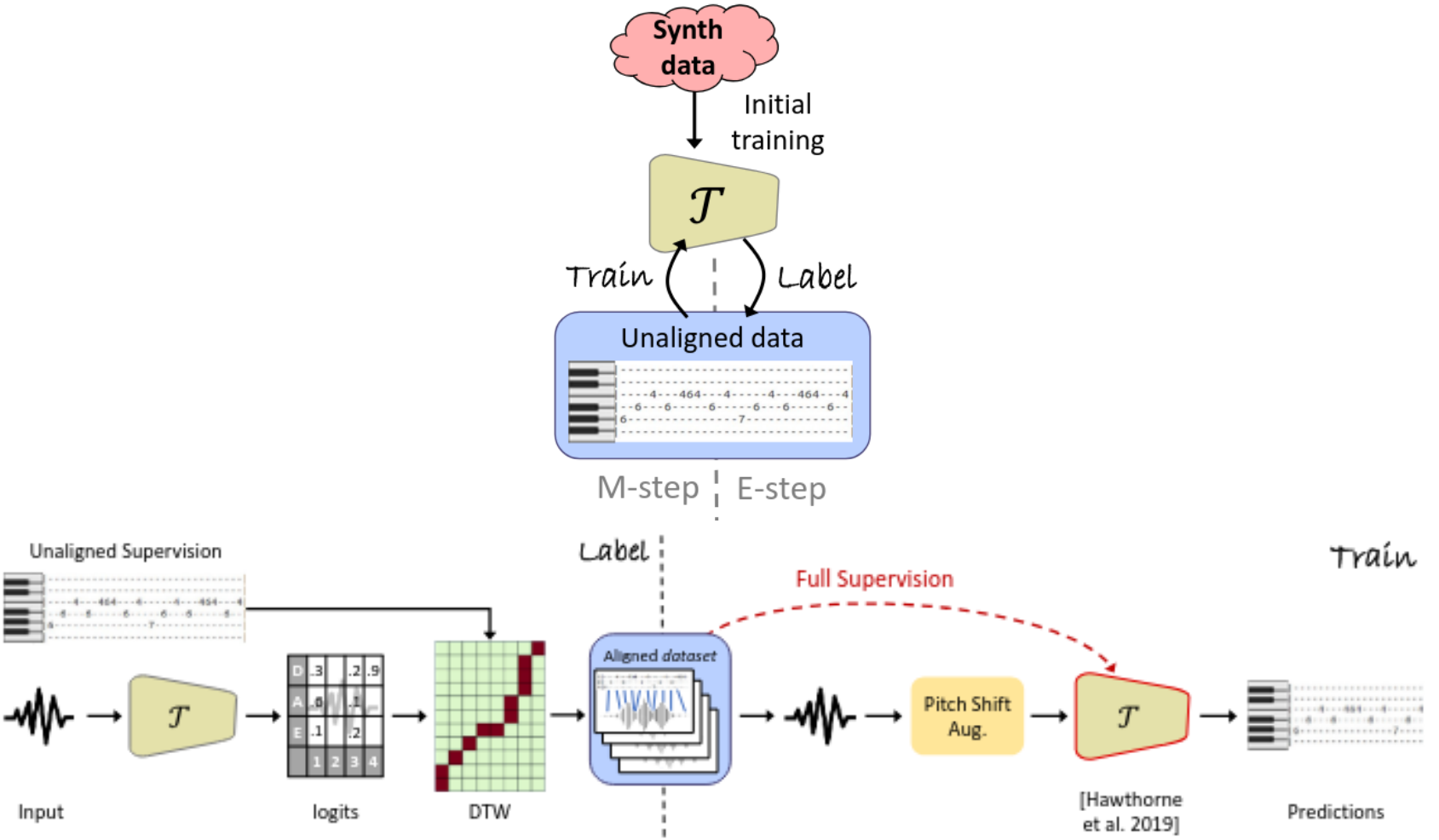

Figure 4: Source: [MamanUnalignedAMT].

Unaligned Supervision for AMT in the Wild II

Dynamic Time Warping[3]

- algorithm for measuring **similarity** between two **temporal sequences**, which may vary in speed
- **optimal match** between two given sequences with following rules:
  - *one or more* matches
  - *first index* must match with first index
  - *last index* must match with last index
  - mapping of the indices must be *monotonically increasing*

Unaligned Supervision for AMT in the Wild III

Pitch Shift Augmentation

- 11 additional **pitch shifted copies** of the data, with pitch shifts (in semitones):

$$s_i = i + \alpha_i, \quad -5 \leq i \leq 5, \quad \alpha_i \sim \mathbf{U}(-0.1, 0.1)$$

- labels computed only for original copy, then shifted accordingly
- data augmentation
- enforce consistency across pitch shift $\Rightarrow$ *learn tonality*

Results

Unaligned Supervision

Onsets and Frames Results

[3] [DynamicTimeWarping]

Bibliography I

[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."

Bibliography II

[MelScale] https://en.wikipedia.org/wiki/Mel_scale

[DynamicTimeWarping] https://en.wikipedia.org/wiki/Dynamic_time_warping