

INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

AUDIO
LABS

Winter Term 2026/27

Selected Topics in Deep Learning for
Audio, Speech, and Music Processing

Lecture

Automatic Music Transcription

Stefan Balke, Andreas Brendel, Ben Maman, Meinard Müller

FAU
Friedrich-Alexander-Universität
Erlangen-Nürnberg

Fraunhofer
IIS

Agenda

1. Introduction to music information retrieval (MIR)

2. Data-driven methods for automatic music transcription

3. Score-audio alignment

4. Multi-instrument transcription

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

AUDIO
LABS

Introduction to Music Information Retrieval (MIR)

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

AUDIO
LABS

Music Representations

Waveform

Spectrograms
STFT / Mel /
CQT

MIDI /
Piano Roll

Sheet Music

Continuous Signal

Symbolic

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

AUDIO
LABS

Music Representations

Waveform

Spectrograms
STFT / Mel /
CQT

MIDI /
Piano Roll

Sheet Music

f_0 Contour

Continuous Signal

Symbolic

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

AUDIO
LABS

Levels of Abstraction: Continuous → Semantic

Amplitude samples

Linear frequencies

Perceptual (log) frequencies

Raw Waveform

STFT

Mel / CQT

Pitch

Notes, instruments,
physical time

Musical time

f_0 contour

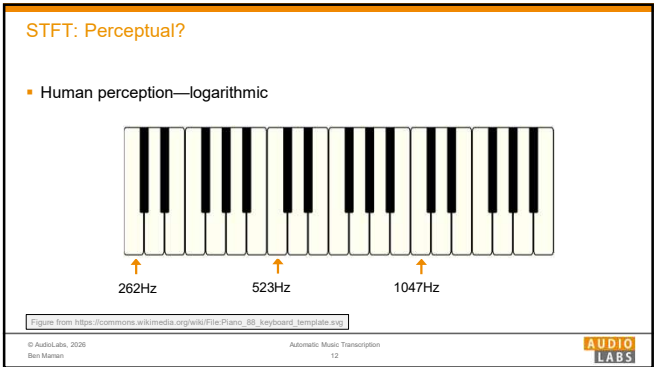
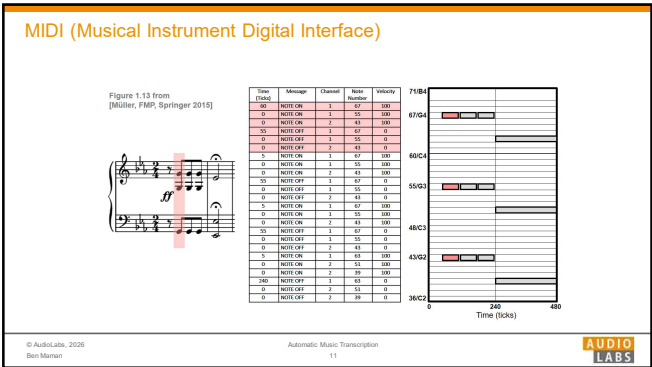
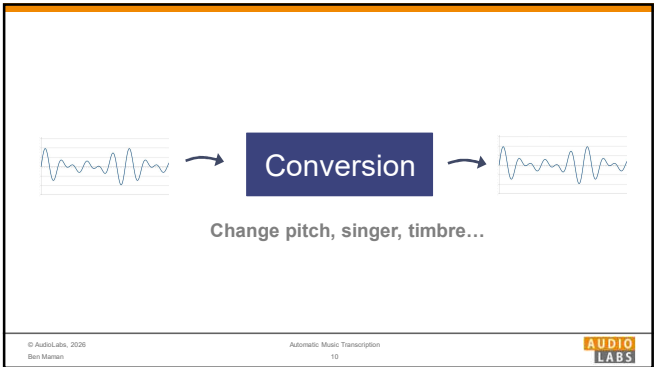
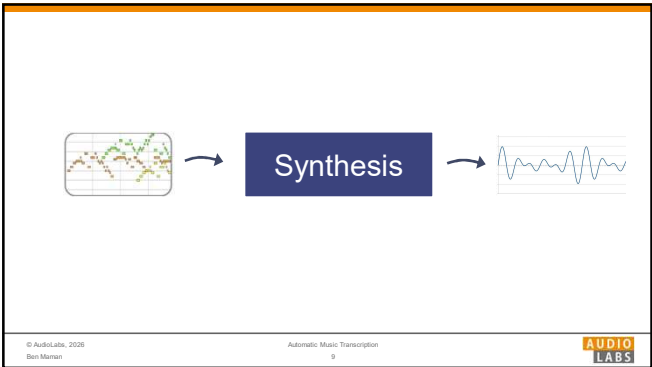
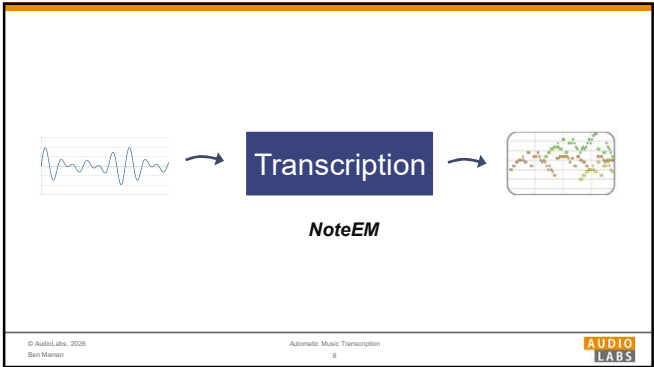
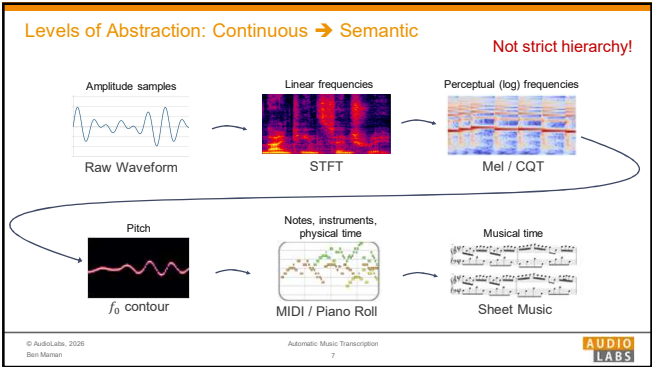
MIDI / Piano Roll

Sheet Music

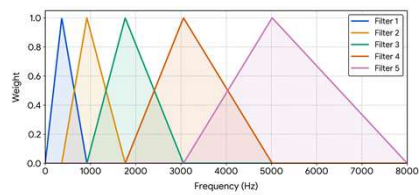
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

AUDIO
LABS



Mel Filterbank



$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right) = 1127 \ln \left(1 + \frac{f}{700} \right)$$

© AudioLabs, 2026

Automatic Music Transcription

13



MIDI Scale

- Perceptual pitch scale, logarithmic
- Relation between frequency in Hertz and MIDI scale:

$$f = 2^{(m-69)/12} \quad m = 12 \log_2(f/440) + 69$$

© AudioLabs, 2026

Automatic Music Transcription

14



Fundamental & Partial Frequencies

- Musical notes - **fundamental** frequencies
- Real-world instruments—many **partial frequencies**
Harmonics & non-harmonics
- Harmonics—integer multiples
- Pitch—**fundamental**
- Timbre—balance between **partials'** strength

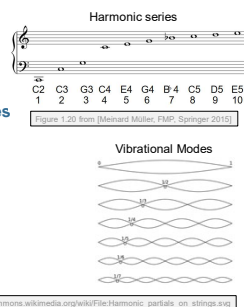


Figure from: commons.wikimedia.org/wiki/File:Harmonicpartials_on_strings.png

© AudioLabs, 2026

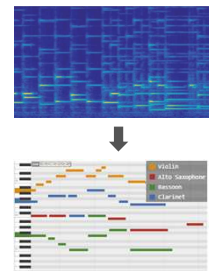
Automatic Music Transcription

15



Music Transcription

- Separate fundamentals from other partials
- "Slicing" - note beginning (onset), end (offset)
- Monophonic (single-voice) → elementary algorithms (no ML)
- Polyphonic & multi-instrument → hard
Mixture of many frequencies
mostly partials, some fundamentals



© AudioLabs, 2026

Automatic Music Transcription

16



Music Synthesis

- Fundamental frequency known
- Concatenative (sampling-based) synthesis
- Challenge – realism
- What makes realistic sound?
Timbre (partial frequencies)
Expressivity—intensity & pitch fluctuation...

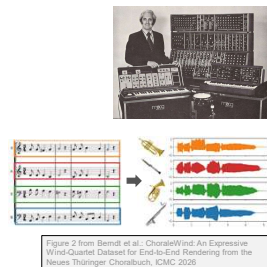


Figure 2 from: Benoit et al., ChoralWind: An Expressive Wind-Quartet Dataset for End-to-End Rendering from the Neural Training Choralbuch, ICAC 2026

© AudioLabs, 2026

Automatic Music Transcription

17



The Analysis–Synthesis Feedback Loop

- How to train analysis model (e.g., transcriber)?
→ Synthesizer: accurately annotated data
- How to learn synthesis from authentic data?
→ Parse with analysis model, e.g., transcriber
Score–audio training pairs



© AudioLabs, 2026

Automatic Music Transcription

18



Automatic Music Transcription

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
18



Automatic Music Transcription: What Is It Good For?

- Music education
- Musicology
- Control signals powering synthesis models

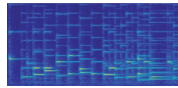
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
20



Automatic Music Transcription

- Input: Audio (spectrogram)
- Output: Note events (MIDI / Piano Roll)



- Challenges:
 - Separate fundamental & partial
 - Timing—note begin / end
 - Polyphony, multi-instrument
 - Instrument detection
 - Dynamics (velocity)

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
21



Automatic Music Transcription

- Single-instrument: e.g., piano
- Instrument-agnostic: detect pitch only within multi-instrument setting
- Instrument-aware: detect pitch & instrument

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
22



Data-Driven Methods for Automatic Music Transcription

© AudioLabs, 2026
Ben Maman

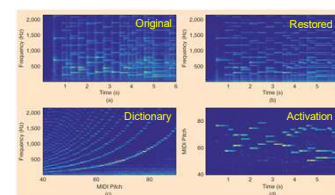
Automatic Music Transcription
23



NMF (Non-Negative Matrix Factorization)

Bermeas et al., Automatic Music Transcription: An Overview, IEEE Signal Processing Magazine, 2019

- $S = DA$ Spectrogram = Dictionary x Activations



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
24



Transcription with Neural Networks

© AudioLabs, 2026
Ben Maman

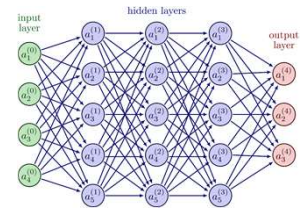
Automatic Music Transcription
25



Neural Network

Composition of:

- Linear maps
- Non-linear activation functions
- Linear maps → **weights**



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
26



Activation Functions

- Make network non-linear → Expressive

Sigmoid
 $\sigma(x) = \frac{1}{1+e^{-x}}$



Leaky ReLU
 $\max(0.1x, x)$



tanh
 $\tanh(x)$



Maxout
 $\max(w_1^T x + b_1, w_2^T x + b_2)$



ReLU
 $\max(0, x)$



ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
27



Loss Functions

- Measure difference between **current** and **target** prediction
- Example: Binary cross-entropy (binary classification)

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

- Objective: Find network weights that minimize loss

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
28



Stochastic Gradient Descent

- Optimization method
- Compute direction of best improvement in loss
- Update network weights

$$w_{t+1} = w_t - \alpha \frac{\partial L}{\partial w_t}$$



- All operations must be differentiable!

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
29



Transcription as Classification

- P pitches, T time frames
- E.g.: P = 88, T = 600
20 seconds, approx. 30 FPS
- Model decides which pitch active at each frame → classification!

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
30



Classification Types

- Monophonic signal → single pitch at a time
- Simple multi-class classification
- Loss: Categorical Cross Entropy
- Polyphonic signal → multiple pitches at a time
- Multi-class, multi-label classification → multiple binary classifications
- Loss: Sum of multiple Binary Cross Entropy

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
31

AUDIO
LABS

Piano Datasets: MAESTRO, MAPS

Hawthorne et al., Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset, ICLR 2019

- Fully supervised datasets
- Audio & MIDI from Disklavier
- Nearly perfect alignment
- MAPS—train synth, test real.
- MAESTRO—130 hours real



© AudioLabs, 2026
Ben Maman

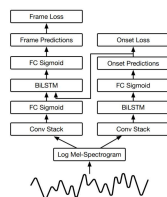
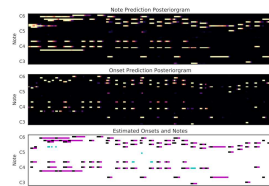
Automatic Music Transcription
32

AUDIO
LABS

Onsets and Frames

Hawthorne et al., Onsets and Frames: Dual-Objective Piano Transcription, ISMIR 2017

- Focus: Note onsets
- Complementary: Note frame-level activity detection



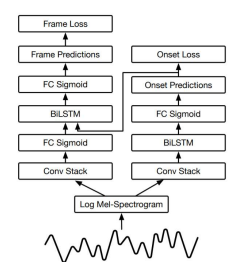
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
33

AUDIO
LABS

Onsets and Frames

- Convolutional layers & long short-term memory (LSTM) layers
- Convolution: Learned pattern-matching filters
- Pitch & time equivariance
- LSTM: Broader temporal context



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
34

AUDIO
LABS

Other Instruments / Genres? Multi-Instrument?

- Only poorly supervised data



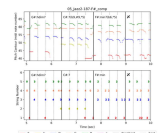
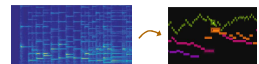
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
35

AUDIO
LABS

Data Problem

- We can't annotate everything
Expensive, sometimes impossible
Sequences — even worse
- Need unsupervised or weakly-supervised learning



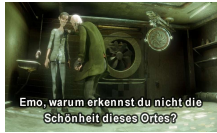
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
36

AUDIO
LABS

Unaligned Supervision

- Full supervision from manual annotation infeasible
Often practically impossible (*precise timing in milliseconds, e.g., music*)
- Unaligned / weak supervision easy**



© AudioLabs, 2026
Ben Maman

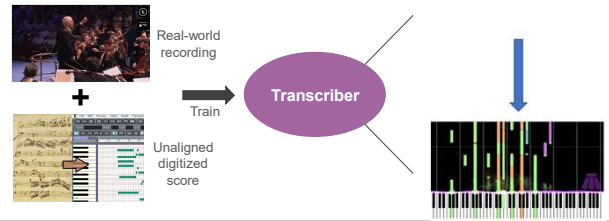
Automatic Music Transcription

37

AUDIO
LABS

Unaligned Supervision

- Create accurate supervision for easily attainable data



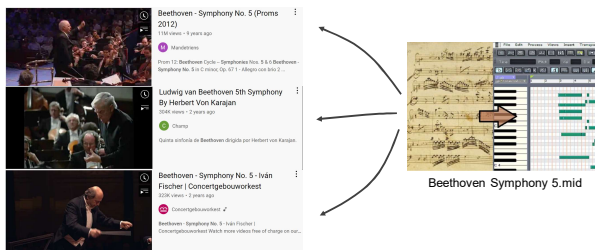
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

38

AUDIO
LABS

Unaligned Supervision



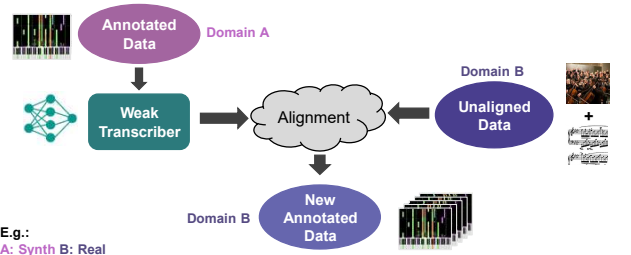
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

39

AUDIO
LABS

Unaligned Supervision



E.g.:
A: Synth B: Real
A: Piano B: Guitar

© AudioLabs, 2026
Ben Maman

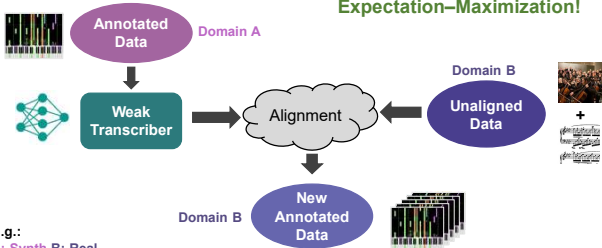
Automatic Music Transcription

40

AUDIO
LABS

Unaligned Supervision

Expectation–Maximization!



E.g.:
A: Synth B: Real
A: Piano B: Guitar

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

41

AUDIO
LABS

Unaligned Supervision

- Key insight: Alignment easier than transcription
- Detect **When** note occurs easier than detecting **whether or not** it occurs
- Weak model can still produce good alignment!

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription

42

AUDIO
LABS

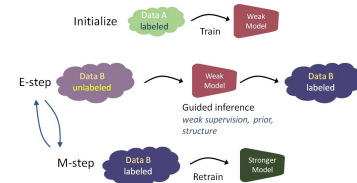
Expectation–Maximization

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
43

AUDIO
LABS

Expectation–Maximization (EM)



- Unsupervised / weakly-supervised learning
- label and train “simultaneously”
- E.g., K-Means, Baum-Welch, Inside-Outside, Deep Clustering

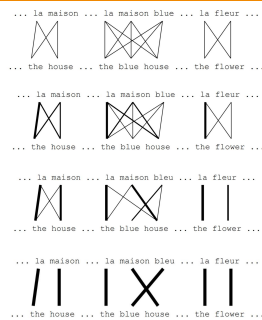
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
44

AUDIO
LABS

EM-Based Machine Translation

- IBM Alignment Model
- Learn word-level translation
Haus => house, building, home...
- Only paired sentences
- Weakly-supervised



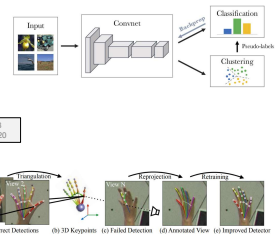
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
45

AUDIO
LABS

Self-Refinement

- Can a network annotate data for itself?
 - Naïve approach—pseudo labels
 - Noisy—perpetuate mistakes
 - Can still improve
- Caron et al., Deep Clustering for Unsupervised Learning of Visual Features, ICCV 2018
Xie et al., Self-training with Noisy Student improves ImageNet Classification, CVPR 2020



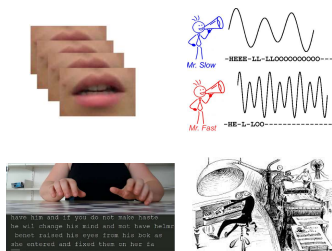
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
46

AUDIO
LABS

Continuous signal → Natural Language

- Speech recognition
- Lip reading
- Flat-surface text entry
- Automatic music transcription
- Unaligned labels often available
- Good prior for EM



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
47

AUDIO
LABS

Score–Audio Alignment

© AudioLabs, 2026
Ben Maman

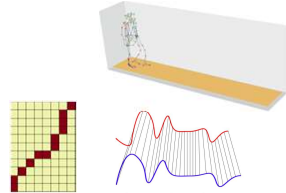
Automatic Music Transcription
48

AUDIO
LABS

Dynamic Time Warping (DTW)

- Optimal monotonic warping between two sequences
- Our case: Warp score to audio
- Dynamic programming solution
- Need local cost function

$$\begin{aligned} &\text{minimize}_{t \in \mathbb{Z}^n} \sum_{i=1}^n C(\mathbf{X}_{t_i}, \mathbf{Y}_i) \\ &\text{subject to} \quad t_0 = 0, \\ &\quad t_n = m, \\ &\quad t_i \leq t_j \quad \text{if } i < j. \end{aligned}$$



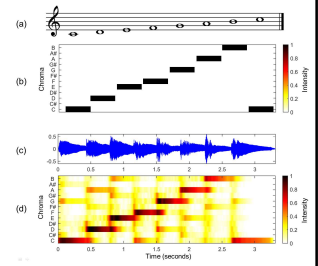
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
49

AUDIO
LABS

DTW: Local Cost Function

- Compare sequence elements, e.g., audio & score
- Audio frame & musical score frame—how similar?
 - Low-frequency STFT features
 - Chroma Features
 - Transcription features



Müller et al., Audio Matching via Chroma-Based Statistical Features, ISMIR 2005
Thickstar et al., Learning Features of Music from Scratch, ICJR 2016
Mannan and Bermann, Unaligned Supervision for Automatic Music Transcription in The Wild, ICML 2022

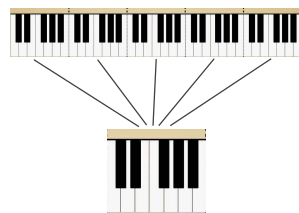
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
50

AUDIO
LABS

Chroma Features

- Maximize activation across octaves
- DTW quadratic time
- 88 pitches → slow
- 12 chroma classes → fast



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
51

AUDIO
LABS

DTW in Polyphonic Settings: Problem

- DTW only sequence-level, not individual notes
- Chord from score not exactly simultaneous in audio? E.g., arpeggio
- Pitch activity detection: Most of note correct ✓
- Onset detection: Many wrong onsets ✗
- Offset detection: Same ✗

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
52

AUDIO
LABS

Sequence- vs. Note-Alignment

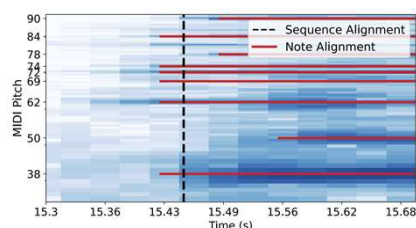


Figure 5 from Berendse et al., Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription

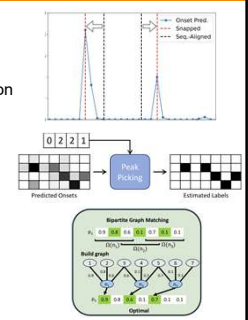
© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
53

AUDIO
LABS

Onset "Snapping"

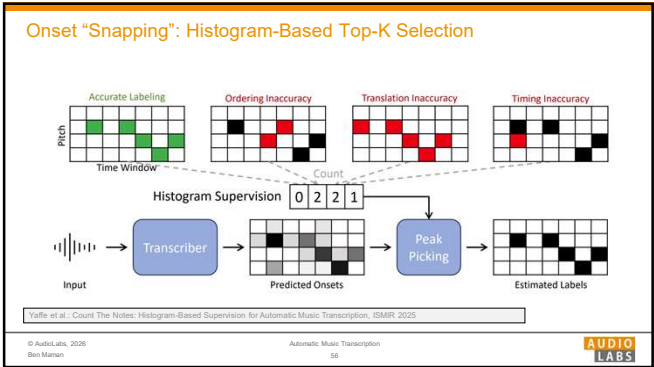
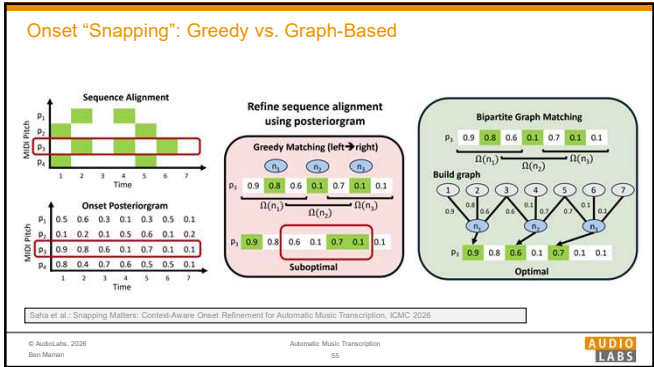
- Locally adjust to timesteps with high onset activation
- Onset features from pre-trained model
- Snapping strategies
 - Histogram-based, greedy nearest-peak, graph-based
- DTW—improve localization



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
54

AUDIO
LABS



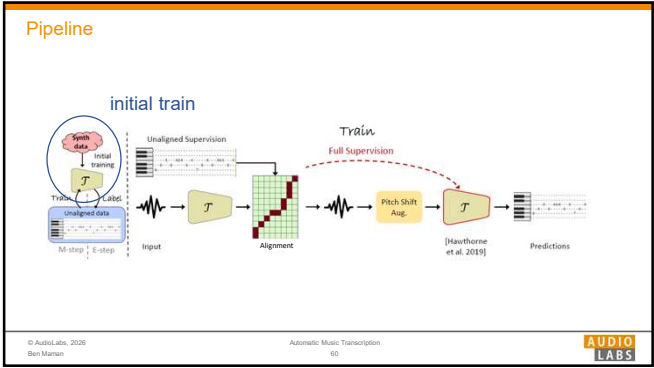
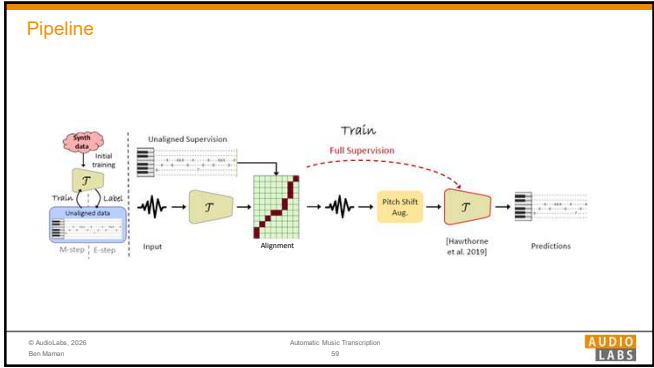
Note Duration, Note Offsets

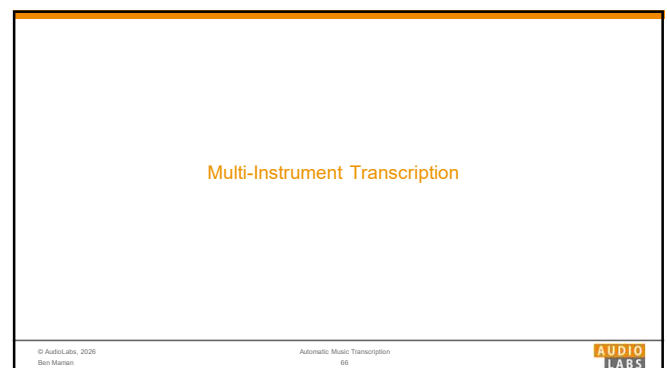
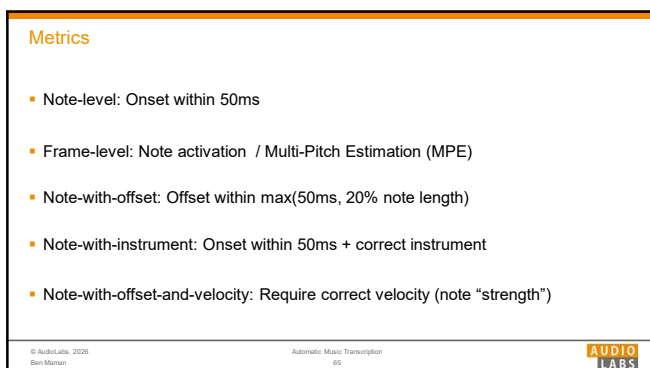
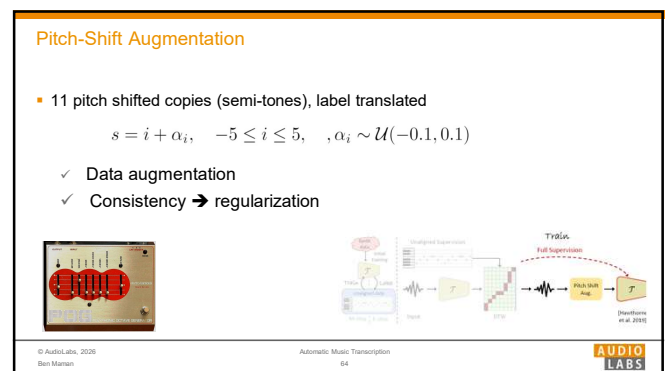
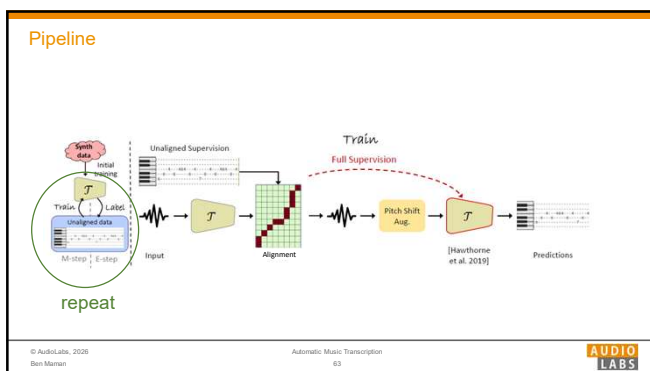
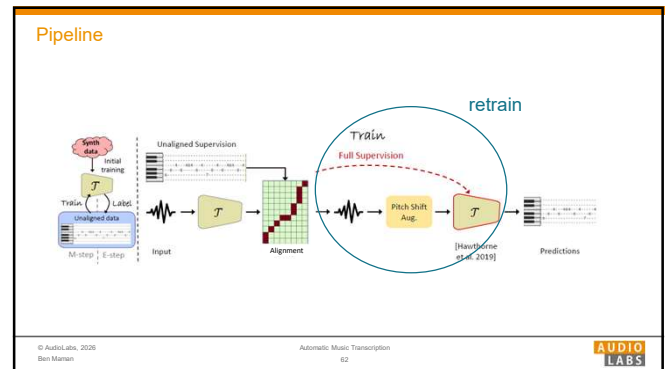
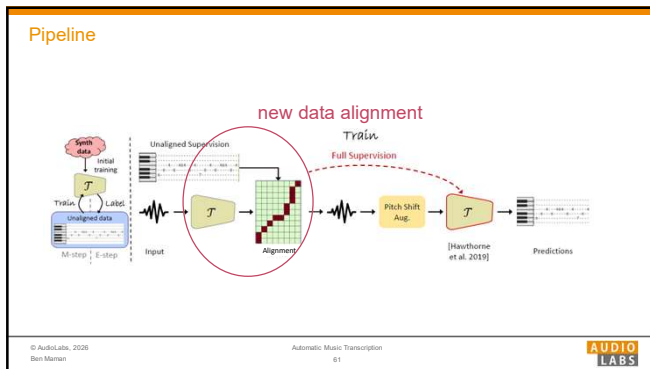
- Remains hard problem!
- Score: only approximate duration
- Offsets hard to detect, even hard to define
- Partial solutions:
 - Synthetic supervised data
 - Real data, approximate duration from score
 - Offset snapping?

© AudioLabs, 2020
Ben Maman
Automatic Music Transcription
57
AUDIO LABS

NoteEM

© AudioLabs, 2020
Ben Maman
Automatic Music Transcription
58
AUDIO LABS





Instrument-Aware Transcription

- Task: Detect pitch & instrument
- Joint modelling task $\rightarrow$ complex
- P pitches, I instruments $\rightarrow P \times I$ classes!



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
67

AUDIO
LABS

Multi-Instrument Transcription

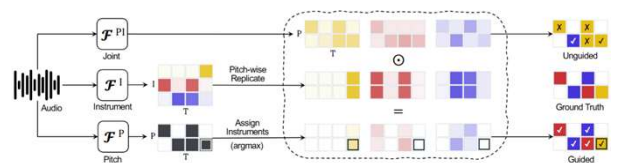
- Instrument-agnostic transcription: Model $\mathcal{F}^P$ predicts $\hat{M}^P \in [0, 1]^{T \times P}$
- Instrument detection: Model $\mathcal{F}^I$ predicts $\hat{M}^I \in [0, 1]^{T \times I}$
- Instrument-aware transcription: Model $\mathcal{F}^{PI}$ predicts $\hat{M}^{PI} \in [0, 1]^{T \times P \times I}$

© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
68

AUDIO
LABS

Pitch & Instrument Guidance

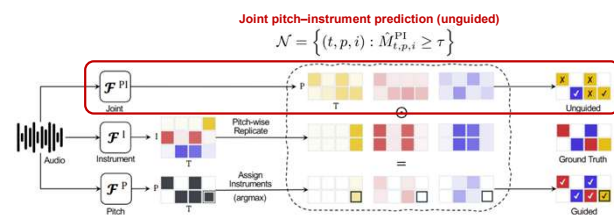


© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
69

AUDIO
LABS

Pitch & Instrument Guidance

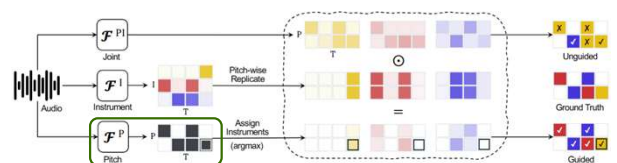


© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
70

AUDIO
LABS

Pitch & Instrument Guidance

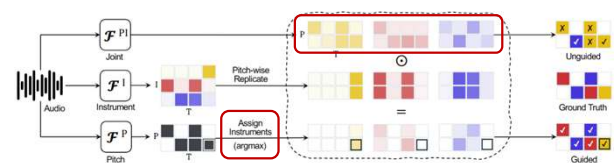


© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
71

AUDIO
LABS

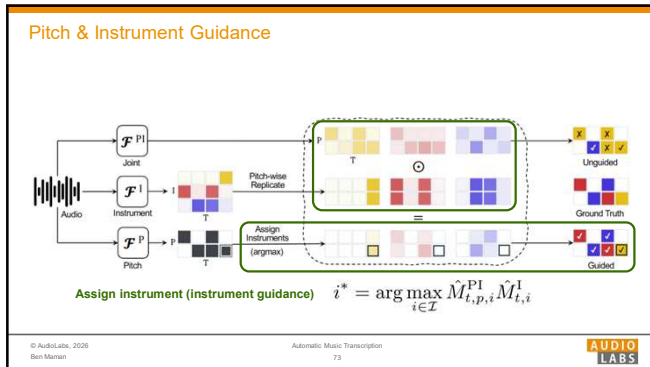
Pitch & Instrument Guidance



© AudioLabs, 2026
Ben Maman

Automatic Music Transcription
72

AUDIO
LABS



Pitch & Instrument Guidance

- Joint pitch-instrument prediction: $\mathcal{N} = \{(t, p, i) : \hat{M}_{t,p,i}^{PI} \geq \tau\}$
 τ : Threshold, e.g., 0.5

Pitch-guided approach:

- Step 1: pitch-only prediction $\mathcal{N} = \{(t, p) : \hat{M}_{t,p}^P \geq \tau\}$
- Step 2: assign instrument $i^* = \arg \max_{i \in \mathcal{I}} \hat{M}_{t,p,i}^{PI}$

© AudioLabs, 2026
Ben Maman
Automatic Music Transcription
76
AUDIO LABS

Pitch & Instrument Guidance

- Instrument guidance: mask instrument assignment using $\mathcal{F}^I$

$$i^* = \arg \max_{i \in \mathcal{I}} \hat{M}_{t,p,i}^{PI} \hat{M}_{t,i}^I$$

- Incorporate instrument model prior

© AudioLabs, 2026
Ben Maman
Automatic Music Transcription
75
AUDIO LABS

Pitch & Instrument Guidance

- Alternative approach: mask using both $\mathcal{F}^I$ and $\mathcal{F}^P$

$$\mathcal{N} = \{(t, p, i) : \hat{M}_{t,p}^P \hat{M}_{t,i}^I \hat{M}_{t,p,i}^{PI} \geq \tau\}$$

- In practice, pitch guidance followed by instrument guidance works better
- Pitch—strong & robust cue, generalizes well across environments

© AudioLabs, 2026
Ben Maman
Automatic Music Transcription
76
AUDIO LABS

Pitch-Only → Instrument-Aware

- Duplicate last layer per instrument
- Initialize—all instruments active, labels—instrument-aware

© AudioLabs, 2026
Ben Maman
Automatic Music Transcription
77
AUDIO LABS

Pitch-Only → Instrument-Aware

Thank You!

© AudioLabs, 2026
Ben Maman
Automatic Music Transcription
78
AUDIO LABS