2025 edition

Deep Learning for Music Analysis and Generation

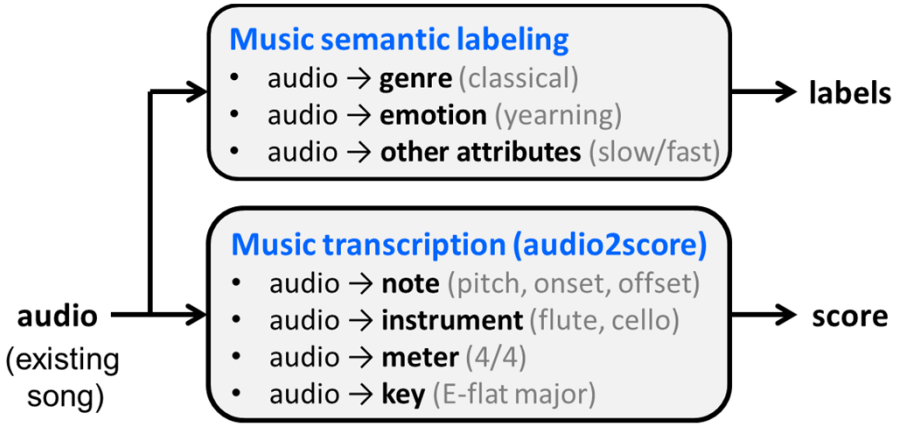
Music Transcription

(audio → scores)

Yi-Hsuan Yang Ph.D.
yhyangtw@ntu.edu.tw

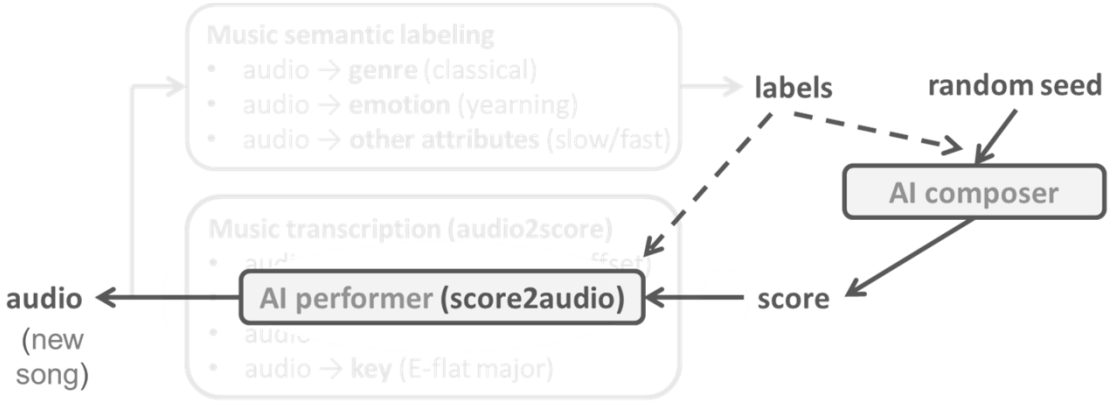
Music AI; or *Music Information Research* (MIR)

- **Music analysis**



  - music understanding
  - music search
  - music recommendation

- **Music generation**



  - MIDI generation
  - audio generation
  - MIDI-to-audio generation

Reference 1: ISMIR 2018 & 2021 Tutorials

https://rachelbittner.weebly.com/other-resources.html

Tutorials

Programming MIR Baselines from Scratch: Three Case Studies 2021

International Society for Music InformationRetrieval (ISMIR) conference

- Part 1: Transcription with NMF (Ethan Manilow)
- Part 2: Pitch Tracking with pytorch (**Rachel Bittner**)
- Part 3: Instrument Classification with OpenL3 & Tensorflow (Mark Cartwright)

📺 See the recording here.

Fundamental Frequency Estimation in Music 2018

International Society for Music Information Retrieval (ISMIR) conference

- Part 1: Pitch (Alain de Cheveigné)
- Part 2: Polyphonic fundamental frequency estimation (**Rachel Bittner**)
- Part 3: Applications (Johana Devaney)

Reference 2: FMP Notebook

https://www.audiolabs-erlangen.de/resources/MIR/FMP/C1/C1.html

Part	Title	Notions, Techniques & Algorithms	HTML	IPYNB
B	Basics	Basic information on Python, Jupyter notebooks, Anaconda package management system, Python environments, visualizations, and other topics	[html]	[ipynb]
0	Overview	Overview of the notebooks (https://www.audiolabs-erlangen.de/FMP)	[html]	[ipynb]
1	Music Representations	Music notation, MIDI, audio signal, waveform, pitch, loudness, timbre	[html]	[ipynb]
2	Fourier Analysis of Signals	Discrete/analog signal, sinusoid, exponential, Fourier transform, Fourier representation, DFT, FFT, STFT	[html]	[ipynb]
3	Music Synchronization	Chroma feature, dynamic programming, dynamic time warping (DTW), alignment, user interface	[html]	[ipynb]

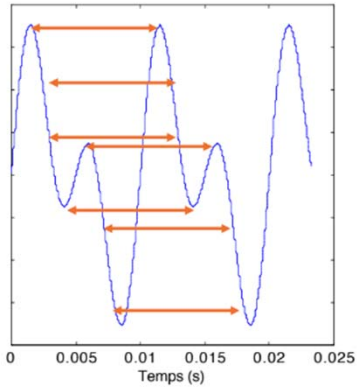
Part	Title	Notions, Techniques & Algorithms	HTML	IPYNB
4	Music Structure Analysis	Similarity matrix, repetition, thumbnail, homogeneity, novelty, evaluation, precision, recall, F-measure, visualization, scape plot	[html]	[ipynb]
5	Chord Recognition	Harmony, music theory, chords, scales, templates, hidden Markov model (HMM), evaluation	[html]	[ipynb]
6	Tempo and Beat Tracking	Onset, novelty, tempo, tempogram, beat, periodicity, Fourier analysis, autocorrelation	[html]	[ipynb]
7	Content-Based Audio Retrieval	Identification, fingerprint, indexing, inverted list, matching, version, cover song	[html]	[ipynb]
8	Musically Informed Audio Decomposition	Harmonic/percussive separation, signal reconstruction, instantaneous frequency, fundamental frequency (F0), trajectory, nonnegative matrix factorization (NMF)	[html]	[ipynb]

Monophonic F0 Estimation

• Part 1: Pitch (Alain de Cheveigné)

- Pitch: depends on fundamental frequency (F0), not on shape
- Frequency-domain F0 analysis
  - problems: harmonics can be stronger, missing fundamental, etc
- Time-domain F0 analysis (based on **auto-correlation)**
  - **YIN** (2002), **pYIN** (2014)

```
librosa.pyin(y, *, fmin, fmax, sr=22050,
```



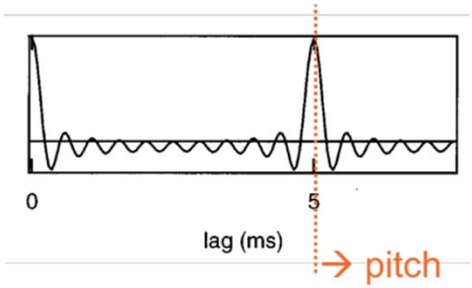
$$A_t(\tau) = \sum_W x(t)x(t-\tau)$$

"shift and compare"

product

time lag

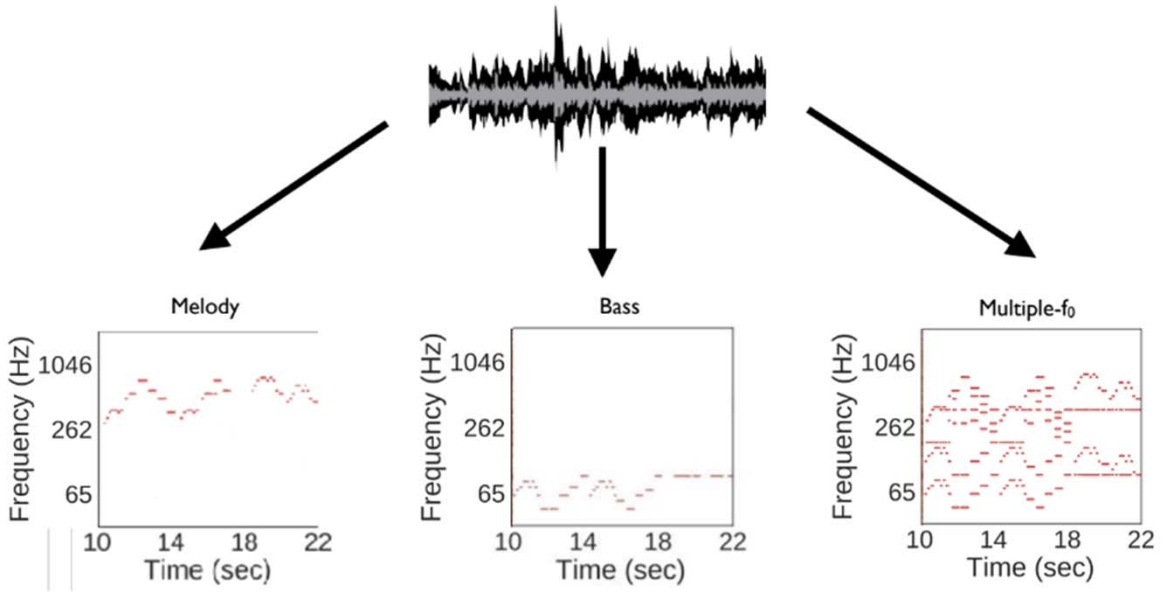
integration window



F0 Estimation

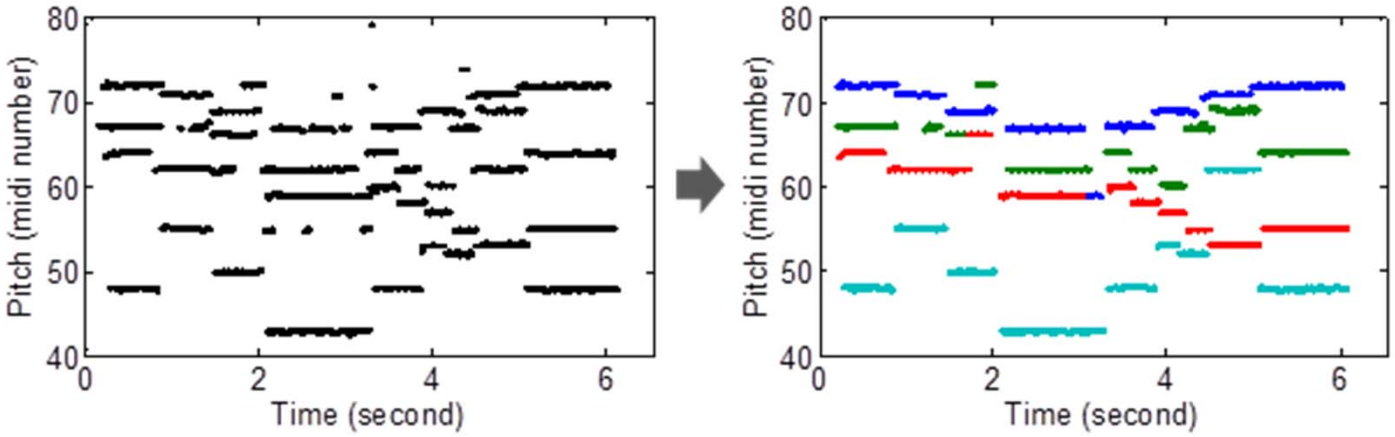
- Part 2: (**Rachel Bittner**)

- Monophonic input vs **polyphonic** input
- Single vs **multiple** F0 estimation
- F0 estimation vs **note** estimation



F0 Estimation: Background

- Related tasks
  - frame-level f0 estimation
  - note estimation
  - streaming
  - score transcription
- Time resolution
- Frequency resolution
- Voicing



Melody Extraction vs. Note Transcription

- **Melody extraction**: F0 (can reflect *overshoot*, *vibrato*, *glissando*, etc)
- **Note transcription**: Note pitch (quantized in frequency)

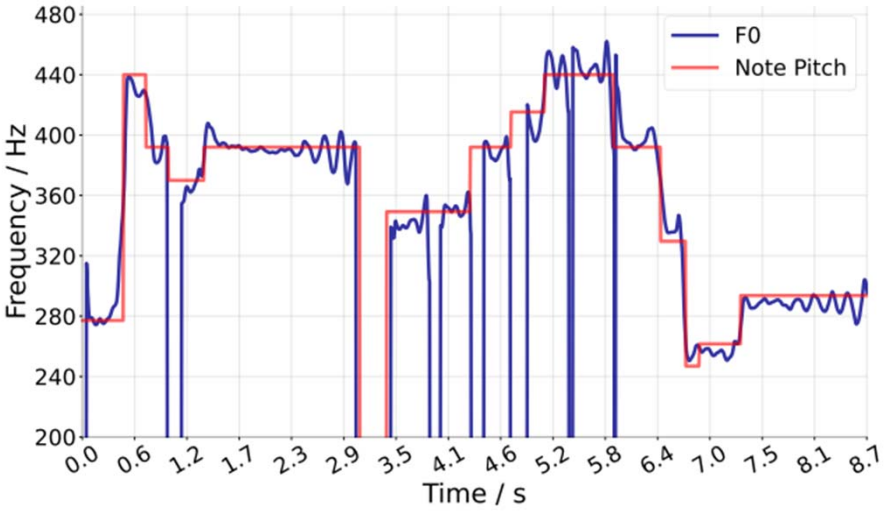


Figure source: M4Singer paper (https://openreview.net/pdf?id=qiDmAaG6mP)

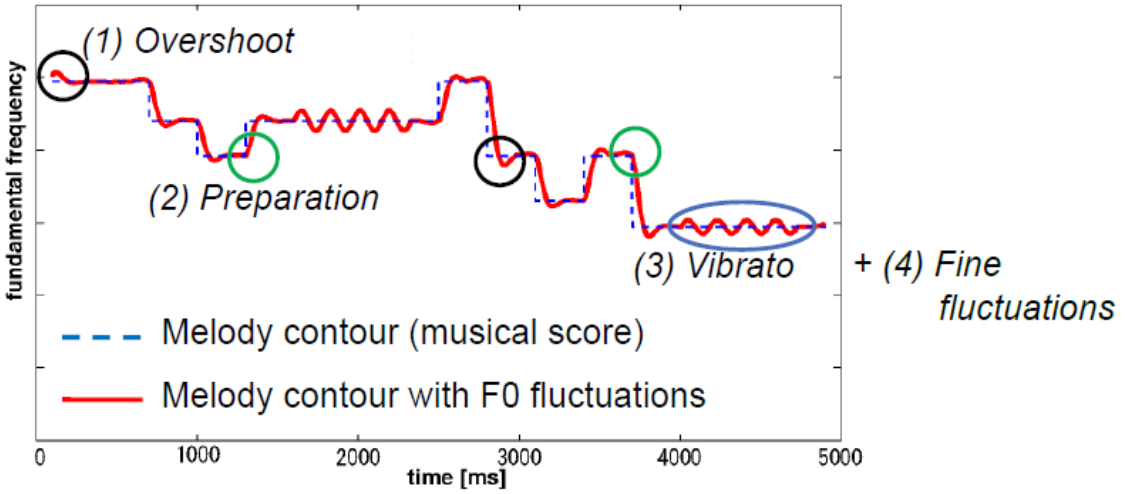
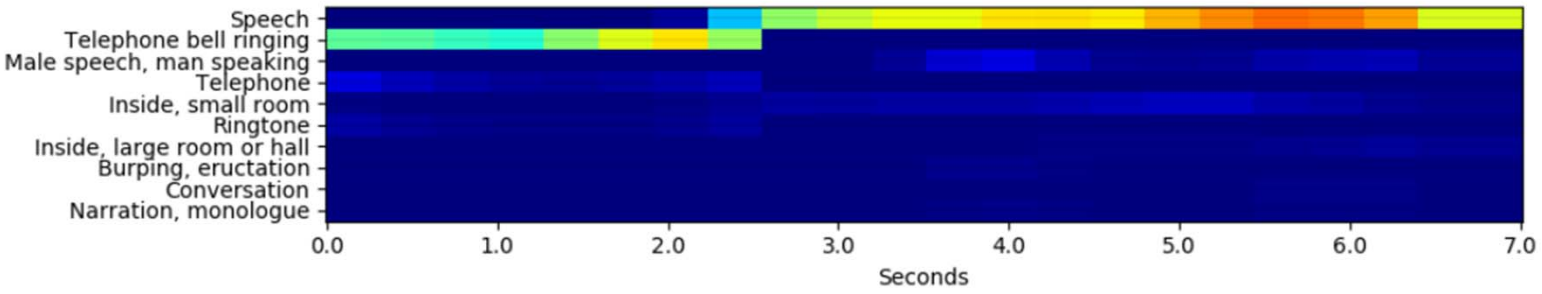
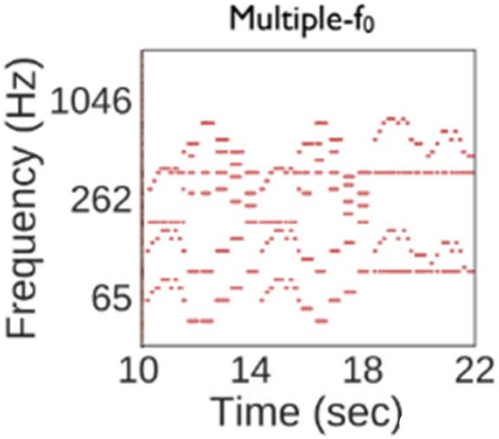


Figure from: Saitou et al, "Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices," WASPAA 2007

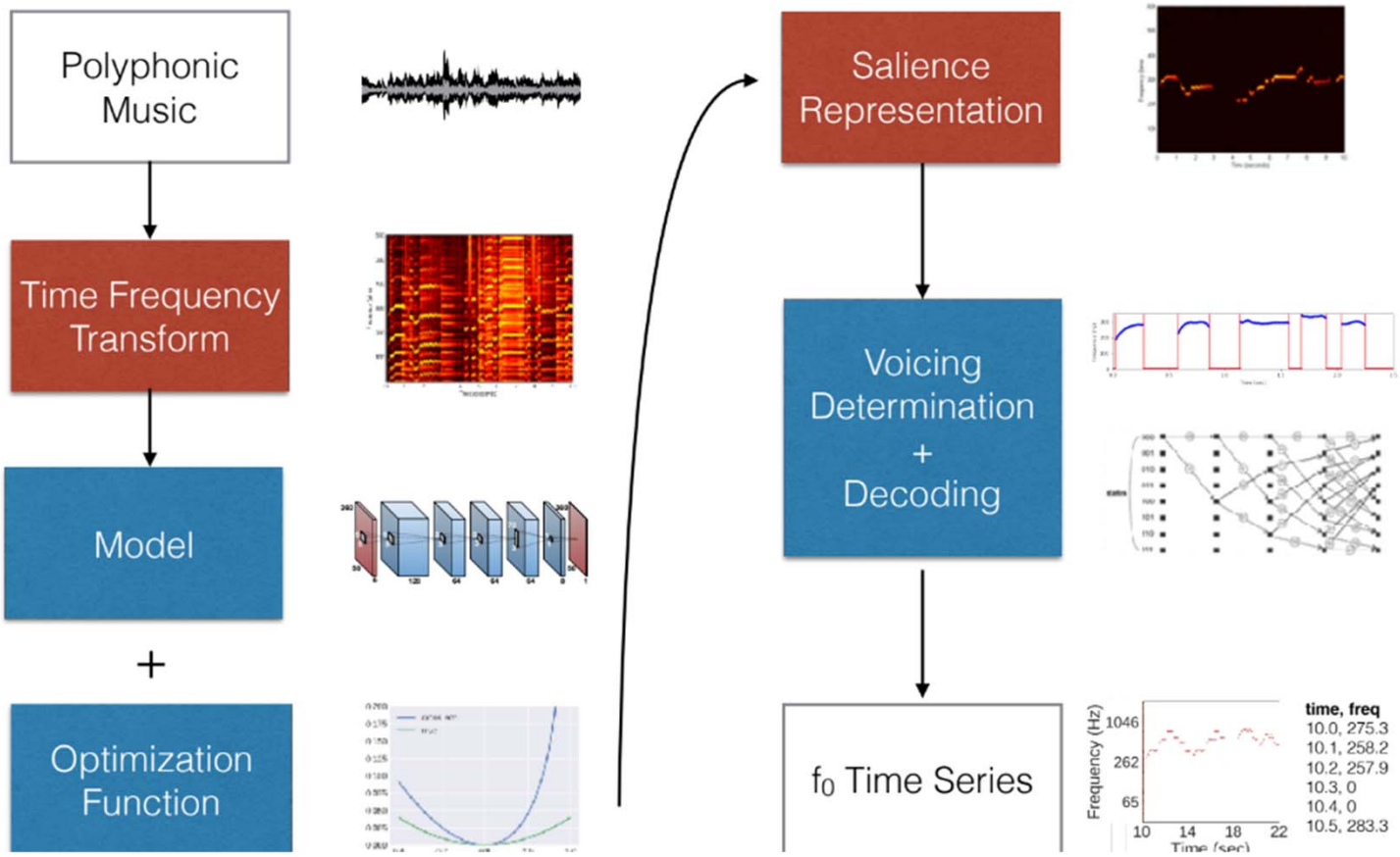
From an ML/DL Viewpoint

- **Per song:** genre classification
  - we can make predictions for each chunk and then aggregate the result over time (e.g., by taking the average logits or by majority voting on the chunk-level decisions)
- **Per short-time chunk:** audio event detection, instrument activity detection
  - e.g., output is a matrix [class x time]
- **Per time-frequency point:** f0 estimation, multi-pitch estimation
  - e.g., output is a matrix [frequency x time]



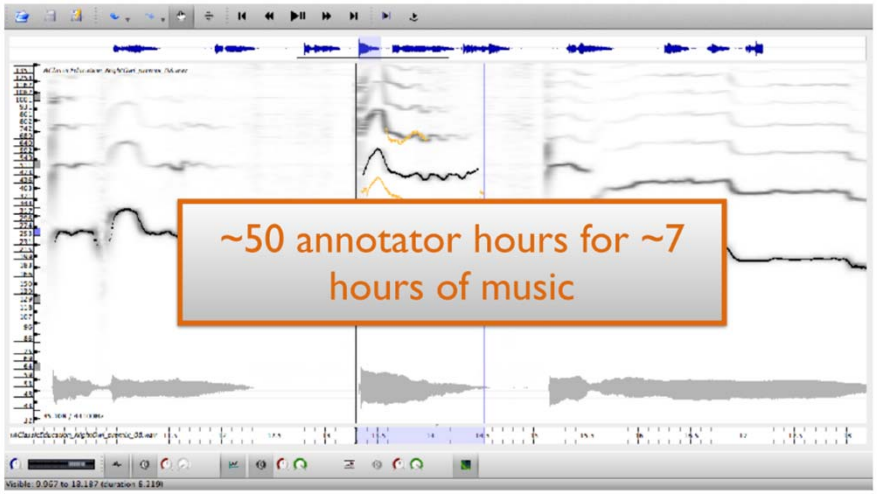


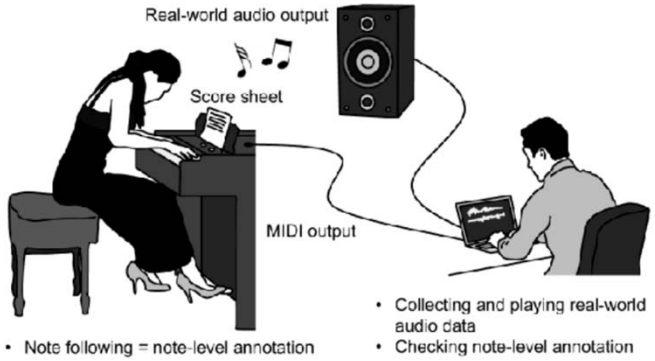
Polyphonic F0 Estimation: Typical Approach



Supervised Training

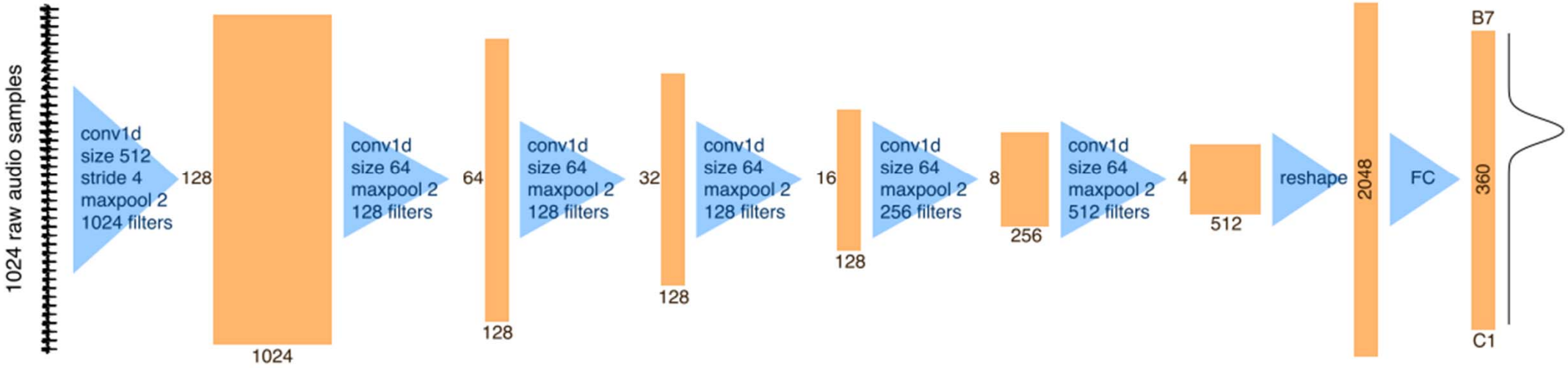
- But labeled data is very hard to collect
- Therefore, progress is a bit slow
  - Single F0 estimation for mono signals is nearly solved
  - Multi-F0 estimation still has room for improvement
  - Note transcription for arbitrary instruments remains hard
    - one exception is piano note transcription, which is nearly solved, partly thanks to high-quality piano synthesizers





Exemplar Model: CREPE (for Monophonic F0 Estimation)

https://github.com/marl/crepe



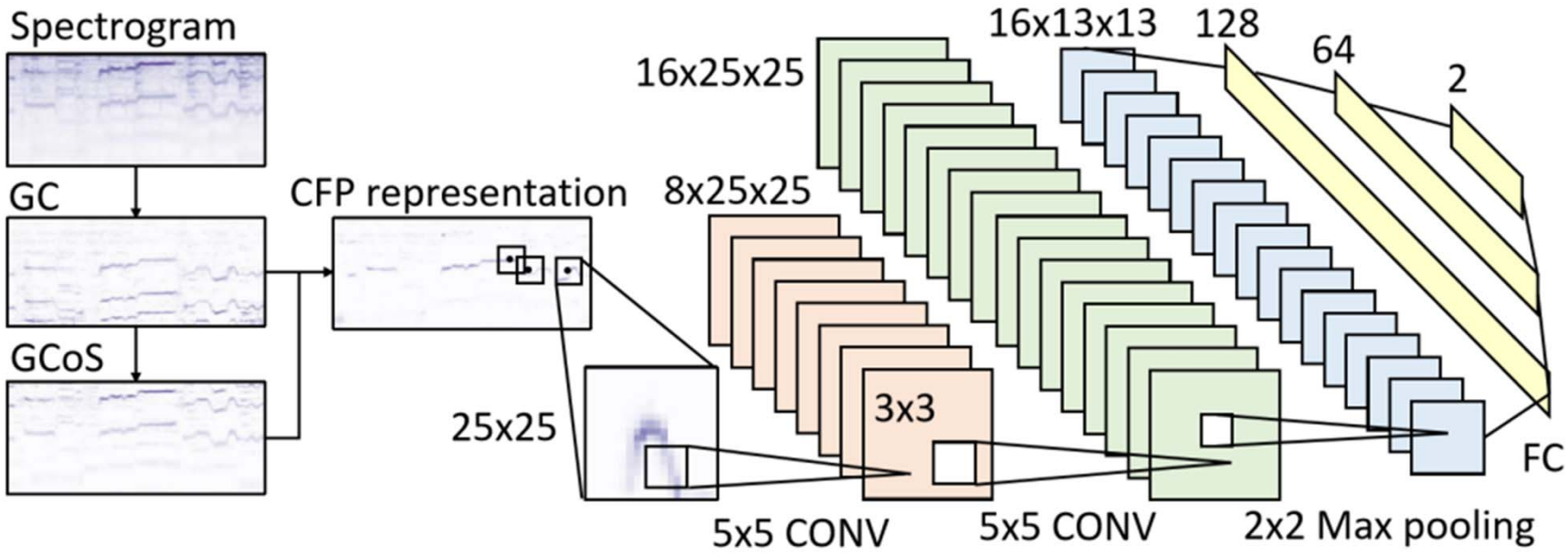
- Make a prediction every 1,024 samples
- Freq resolution: 20 cents (1/5 semitone)
- Outperform DSP-based methods such as pYIN and SWIPE

Dataset	Threshold	CREPE	pYIN	SWIPE
RWC-synth	50 cents	**0.999±0.002**	0.990±0.006	0.963±0.023
	25 cents	**0.999±0.003**	0.972±0.012	0.949±0.026
	10 cents	**0.995±0.004**	0.908±0.032	0.833±0.055

Kim et al., "CREPE: A convolutional representation for pitch estimation," ICASSP 2018

Exemplar Model: ExtPatchCNN (for Monophonic F0 Estimation)

https://github.com/leo-so/VocalMelodyExtPatchCNN

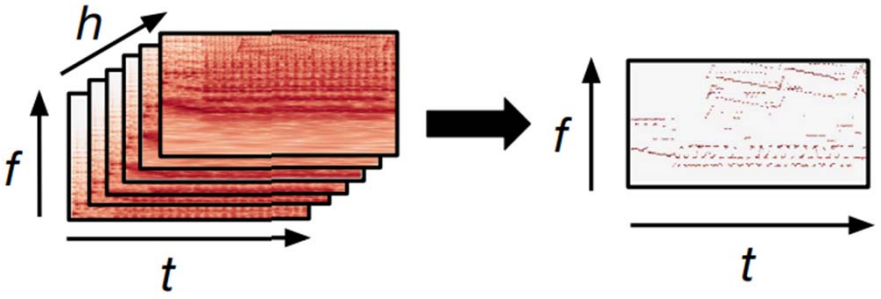


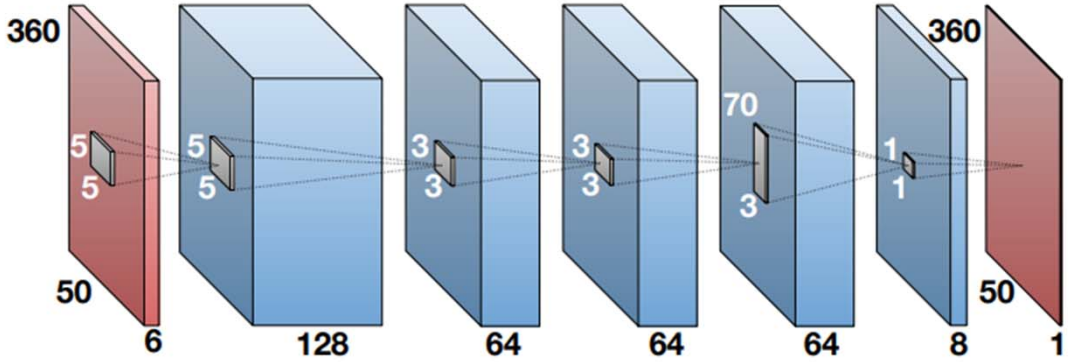
- Binary classification: melody vs non-melody

Su, "Vocal melody extraction using patch-based CNN," ICASSP 2018

Exemplar Model: DeepSalience (for Polyphonic F0 Estimation)

https://github.com/rabitt/ismir2017-deepsalience



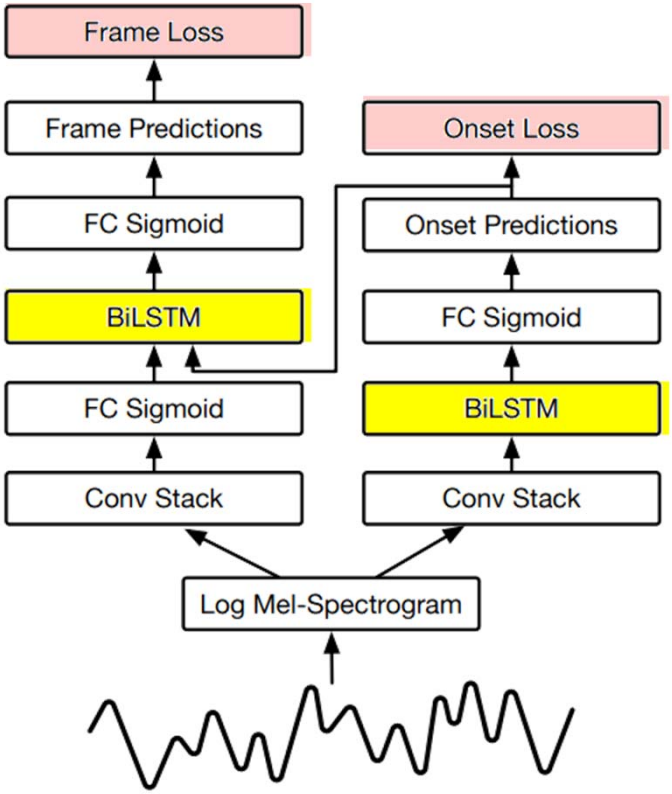


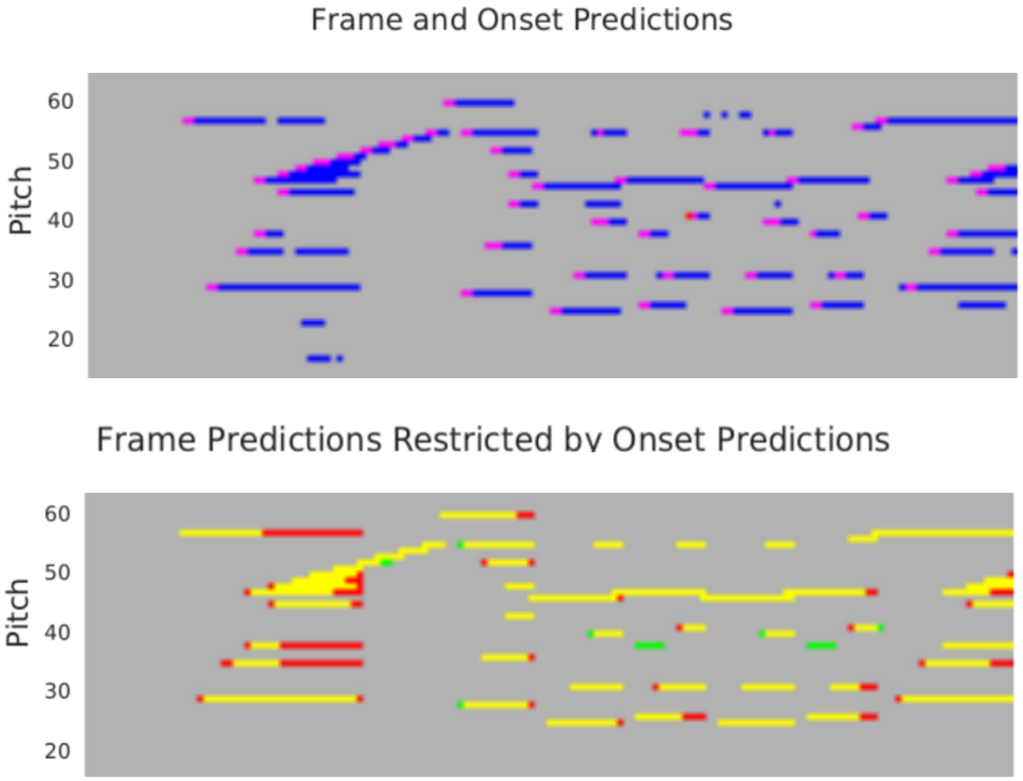
- Input: harmonic constant-Q transform (HCQT) with 6 channels
- Model: 5-layer CNN (**no strides, no pooling → no shift-invariance**)
  - (5 x 5): 1 semitone in frequency and 50 ms in time
  - (70 x 3): 14 semitones in frequency to capture relationships between frequency content within an octave

Kim et al., "Deep salience representations for F0 estimation in polyphonic music," ISMIR 2017

Exemplar Model: Onset-and-Frames (for Piano Transcription)

https://magenta.tensorflow.org/onsets-frames



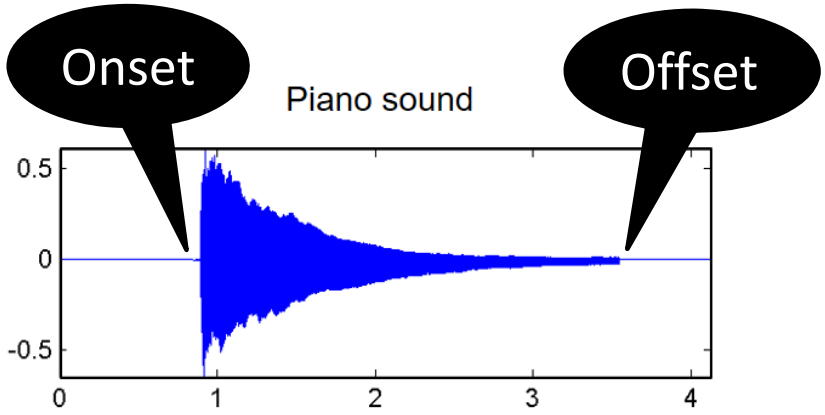


Hawthorne et al., "Onsets and Frames: Dual-objective piano transcription," ISMIR 2018

Pitch, Onset, Offset, Velocity

https://magenta.tensorflow.org/datasets/maestro

"We partnered with organizers of the International **Piano-e-Competition** for the raw data used in this dataset. During each installment of the competition virtuoso pianists perform on **Yamaha Disklaviers** which, in addition to being concert-quality acoustic grand pianos, utilize an integrated high-precision **MIDI capture** and playback system."



Dynamic's note velocity

Dynamic	Velocity*	Voice
ppp	16	Whispering
pp	33	Almost at a whisper
p	49	Softer than speaking voice
mp	64	Speaking voice
mf	80	
f	96	Louder than speaking
ff	112	Speaking loud
fff	126	Yelling

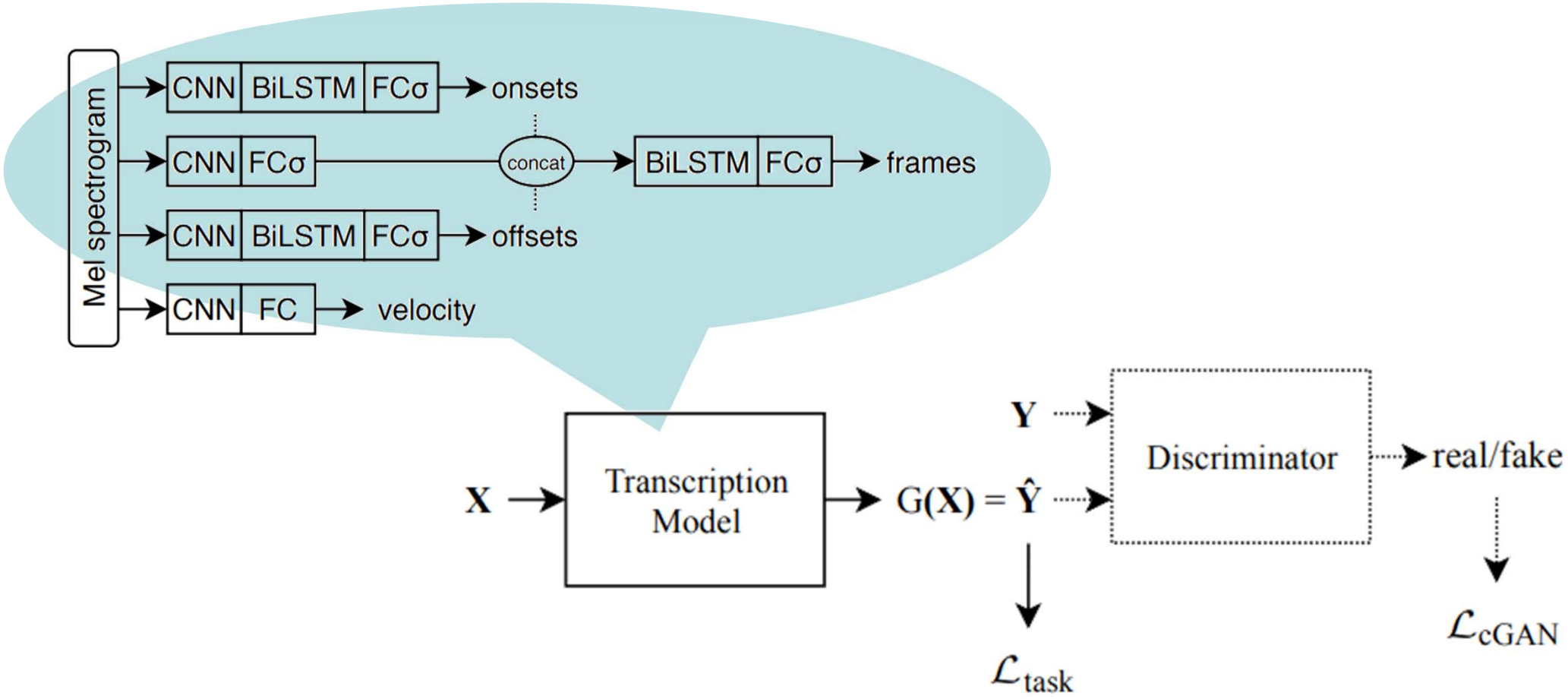
Decrescendo (diminuendo) Crescendo Accent

***Note velocity adopted from Logic Pro**

https://freeonlinesheetmusic.wordpress.com/2016/01/10/dynamics/

Hawthorne et al., "Enabling factorized piano music modeling and generation with the MAESTRO dataset," ICLR 2019

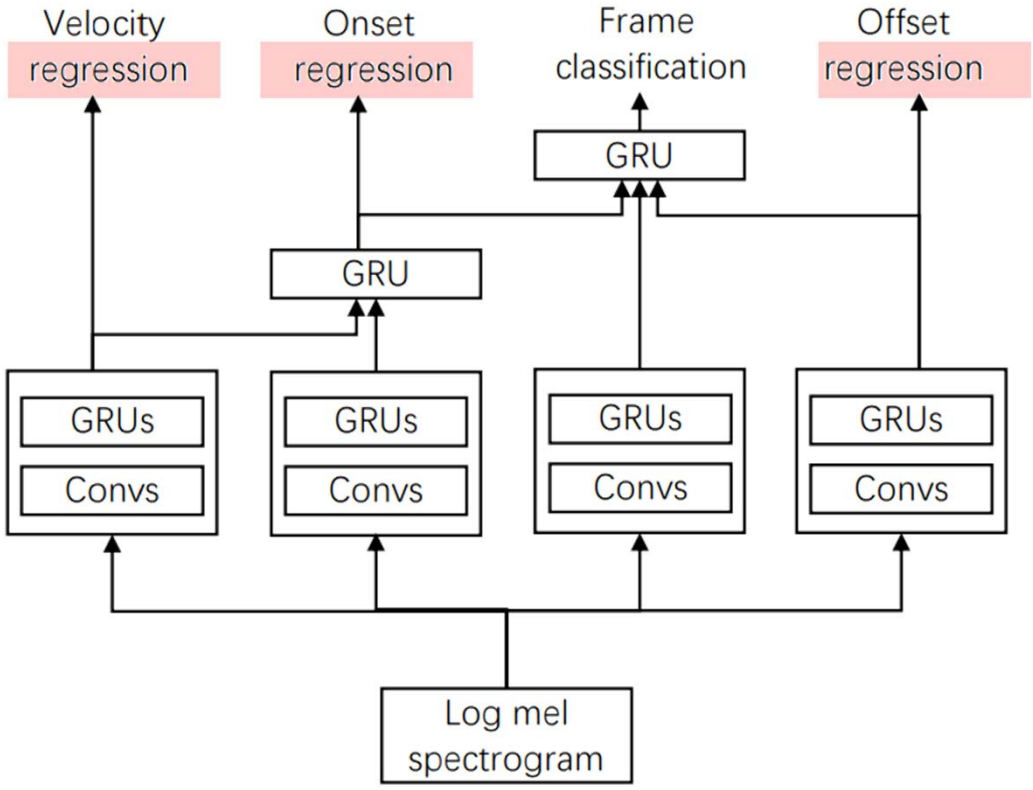
Exemplar Model: Piano Transcription with GAN

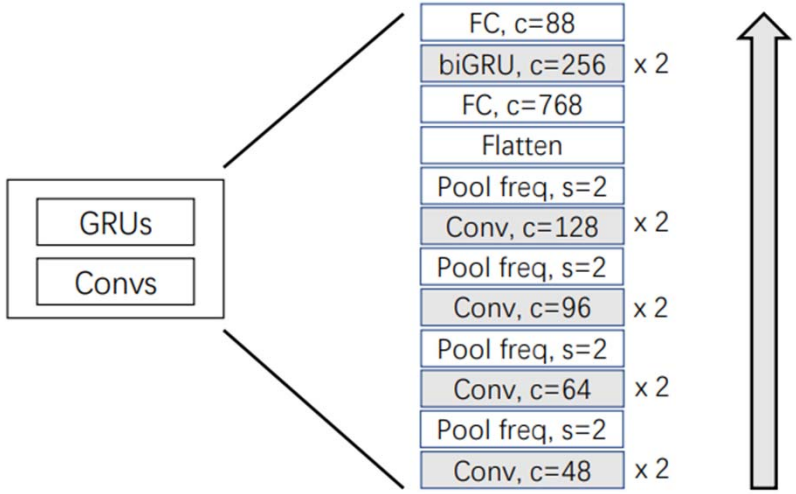


Kim & Bello, “Adversarial learning for improved onsets and frames music transcription,” ISMIR 2019

Exemplar Model: High-resolution Piano Transcription

https://github.com/bytedance/piano_transcription





Kong et al., “High-resolution piano transcription with pedals by regressing onsets and offsets times,” TASLP 2021

Exemplar Model: Basic Pitch (for Note Transcription)

https://github.com/spotify/basic-pitch

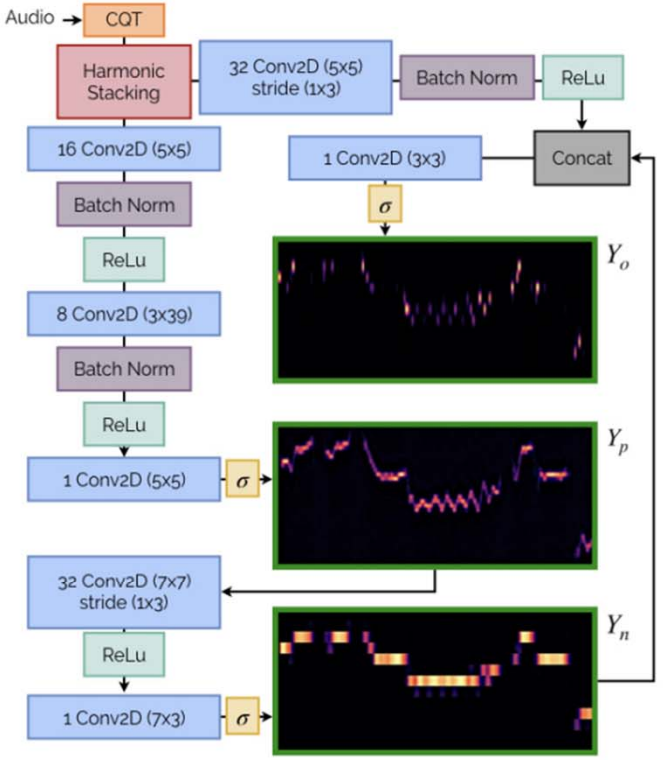
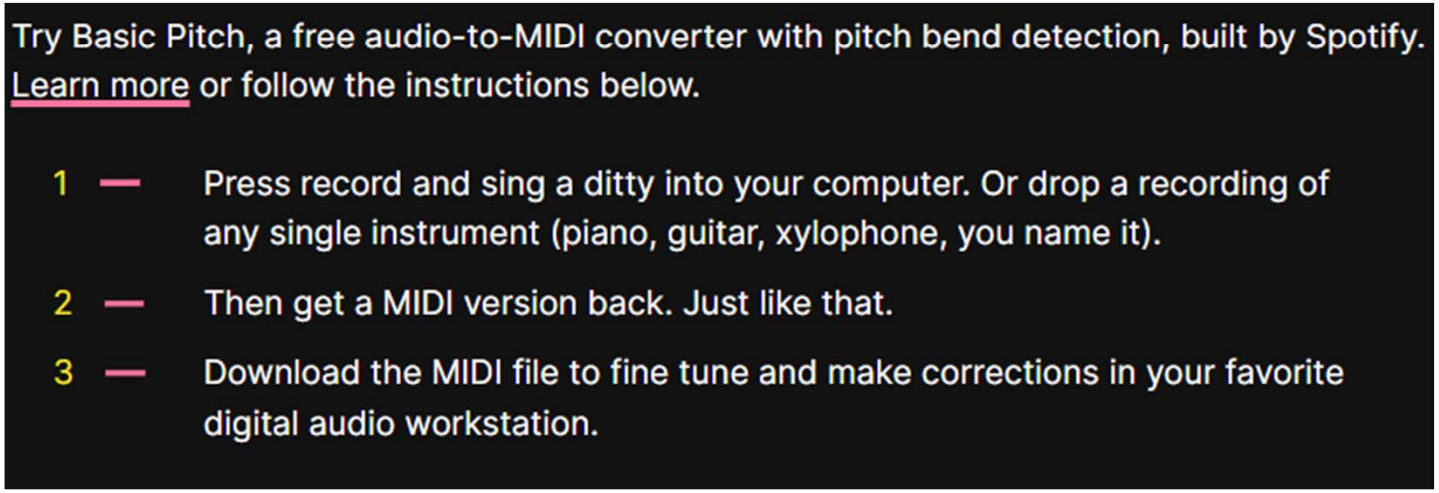


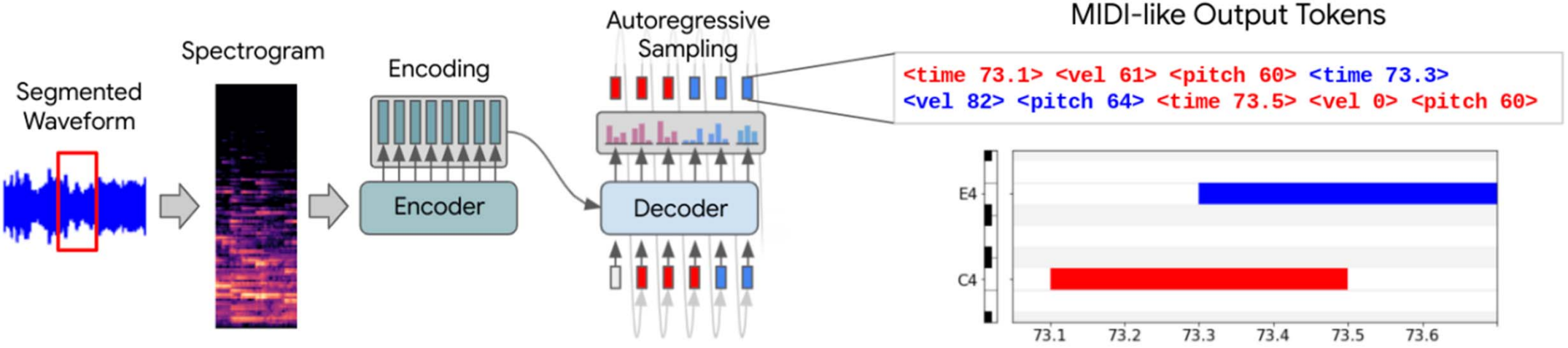
Fig. 1. The NMP architecture. The matrix posteriorgram outputs Y_o, Y_p, and Y_n are outlined in green. σ indicates a sigmoid activation.

- For *instrument-agnostic* note transcription and multi-pitch estimation
- Pure CNN-based for being light-weight



Bittner et al., "A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation," ICASSP 2022

Exemplar Model: Seq2Seq Transformers (for Piano Transcription)



- Jointly modeling audio features and language-like output dependencies
- Possible to pre-train the decoder

	Model	Onset, Offset, & Velocity F1	Onset & Offset F1	Onset F1
MAESTRO V1.0.0	Transformer (ours)	**82.18**	**83.46**	95.95
	Kong et al. 2020 [19]	80.92	82.47	**96.72**
	Kwon et al. 2020 [21]	–	79.36	94.67
	Kim & Bello 2019 [20]	80.20	81.30	95.60
	Hawthorne et al. 2019 [2]	77.54	80.50	95.32

Hawthorne et al., "Sequence-to-sequence piano transcription with Transformers," ISMIR 2021

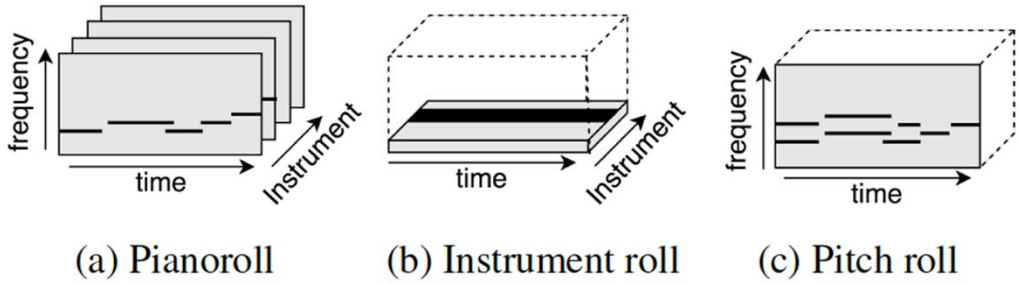
Evaluation Metrics for Pitch-Related Tasks

https://craffel.github.io/mir_eval/

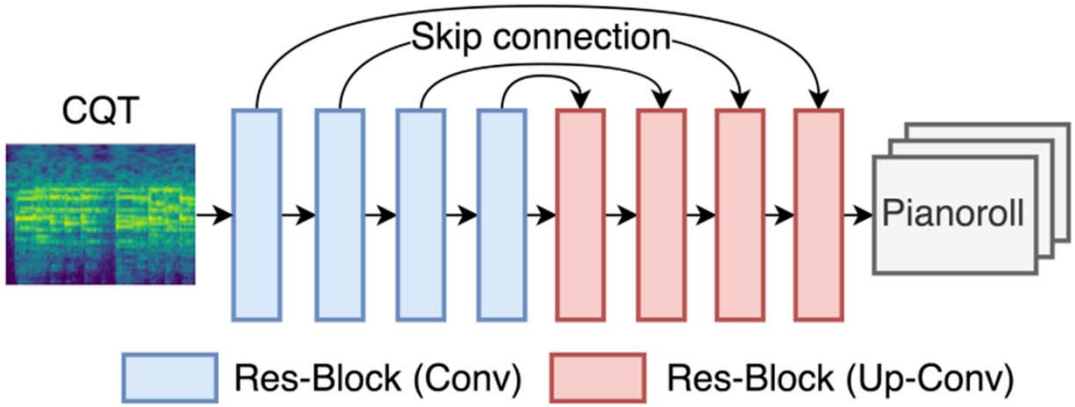
- **Single- and multi-pitch estimation**
  - correct if within 0.5 semitones of a reference frequency
  - chroma accuracy (ignore octave)
  - voicing measures
- **Piano transcription**
  - onset tolerance window: 50ms
  - offset tolerance window: 20% of note duration
  - onset only vs. onset+offset
- **Chord transcription**
  - root
  - majmin
  - majmin_inv
  - thirds
  - triads
  - tetrads
  - sevenths
  - overseg, underseg, seg

Exemplar Model: Joint Prediction of Instrument and Pitch

- Predicting the “instrument roll” (time x instrument)



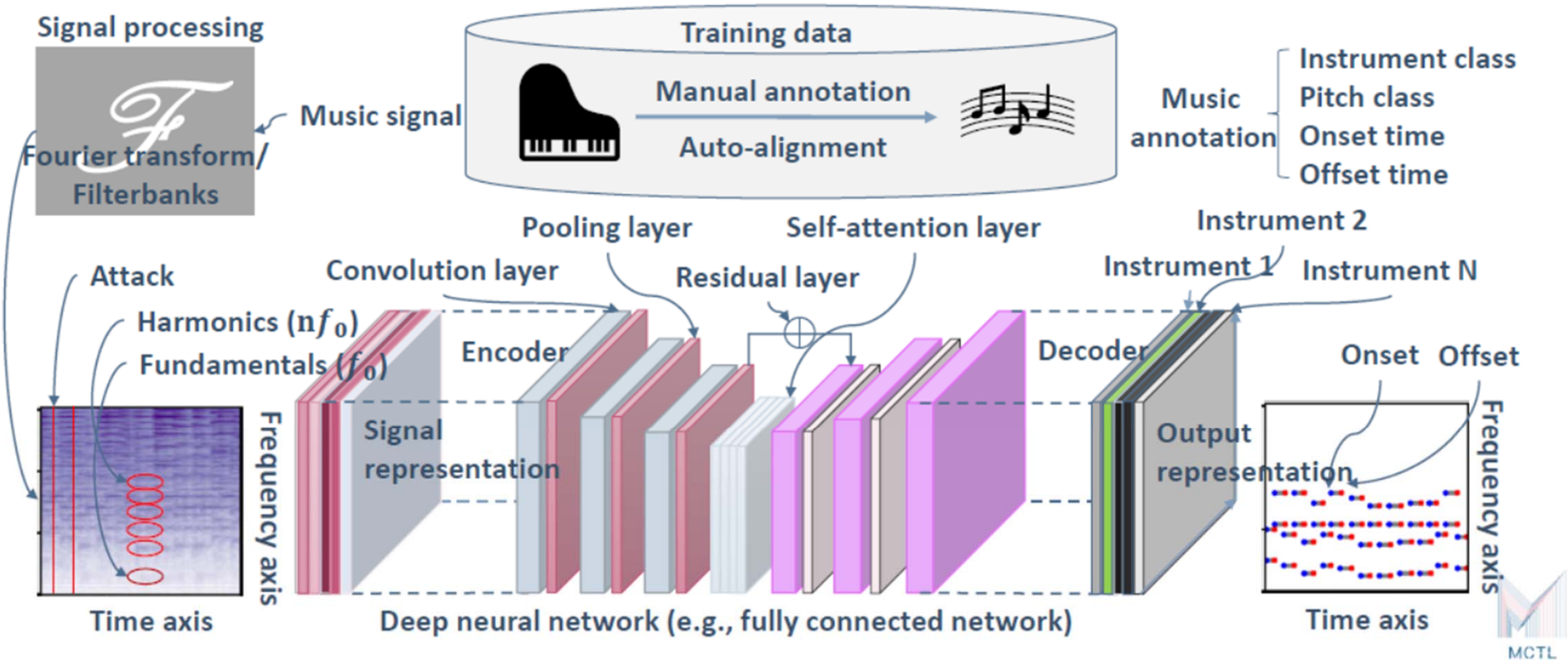
(a) Pianoroll (b) Instrument roll (c) Pitch roll



Hung et al, “Multitask learning for frame-level instrument recognition”, ICASSP 2019

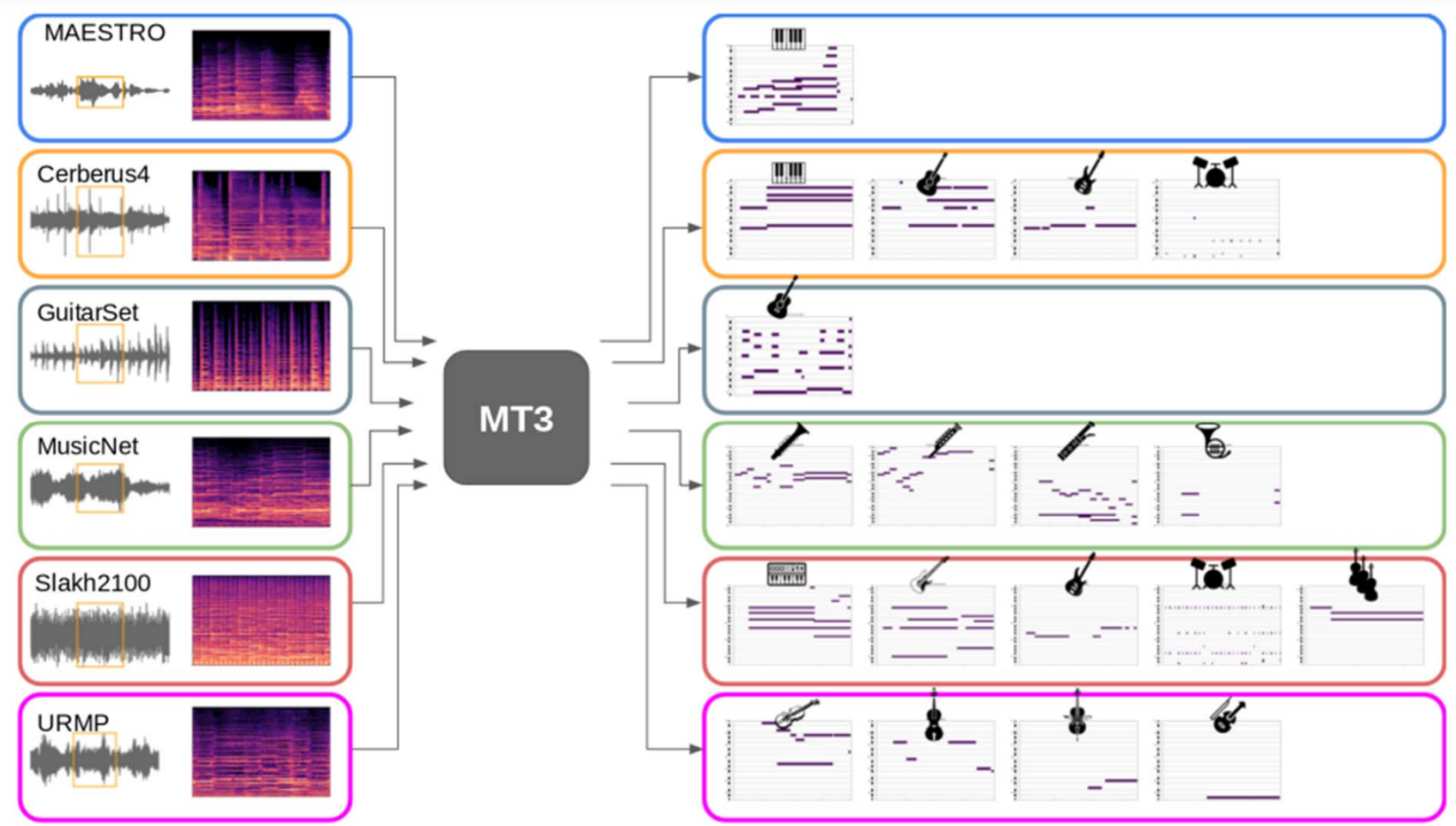
Exemplar Model: Omnizart (for Multi-Instrument Transcription)

https://github.com/Music-and-Culture-Technology-Lab/omnizart



Wu et al., "Omnizart: A general toolbox for automatic music transcription," JOOS 2021

Exemplar Model: MT3 (for Multi-instrument Music Transcription)



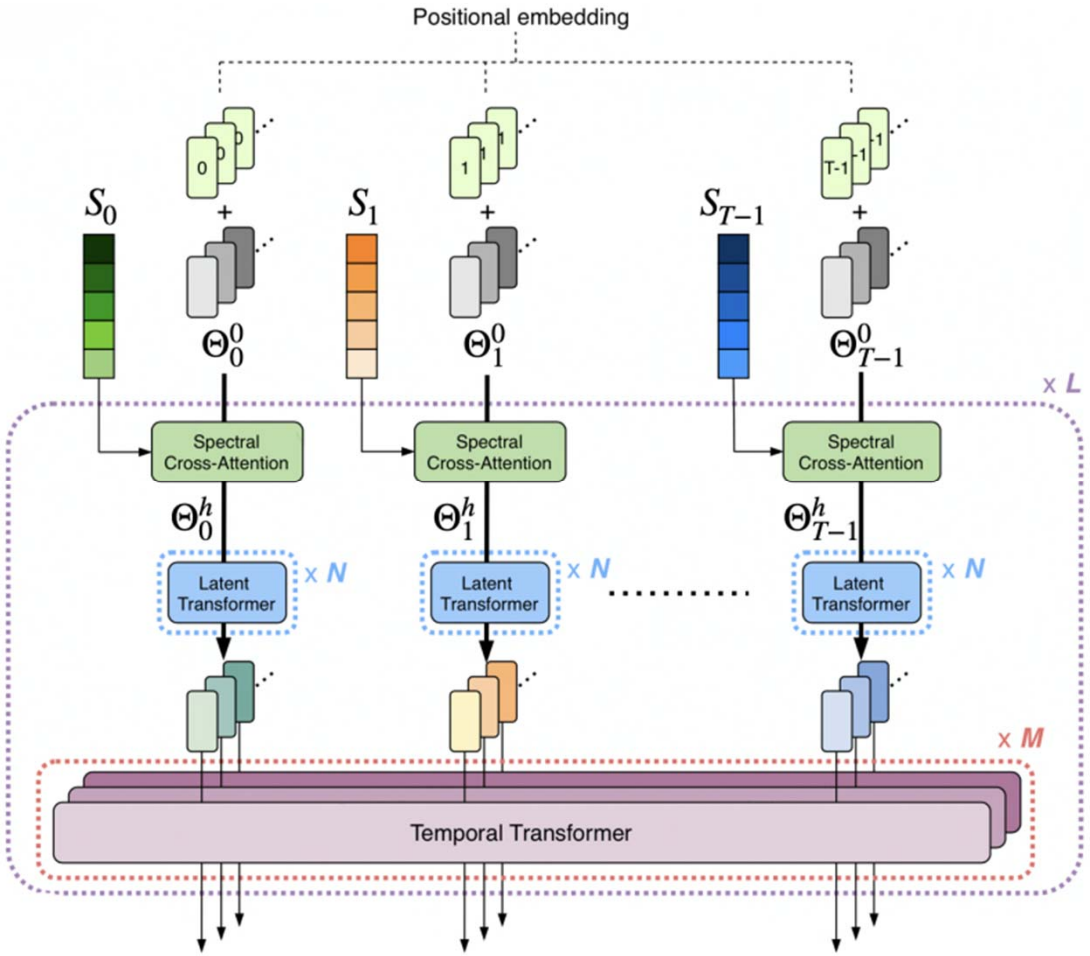
Gardner et al., “MT3: Multi-task multitrack music transcription,” ICLR 2022

Exemplar Model: Perceiver TF (for Multi-instrument Music Transcription)

- Feedforward Transformer-encoder (not encoder/decoder)
- Use audio-domain loss, rather than MT3's symbolic-domain loss
- Onset-and-frames-like loss, but for $J > 1$ (multiple) instruments

$$\mathcal{L} = \sum_{j=0}^{J-1} (l^j_{\text{onset}} + l^j_{\text{frame}})$$

- Outperforms MT3

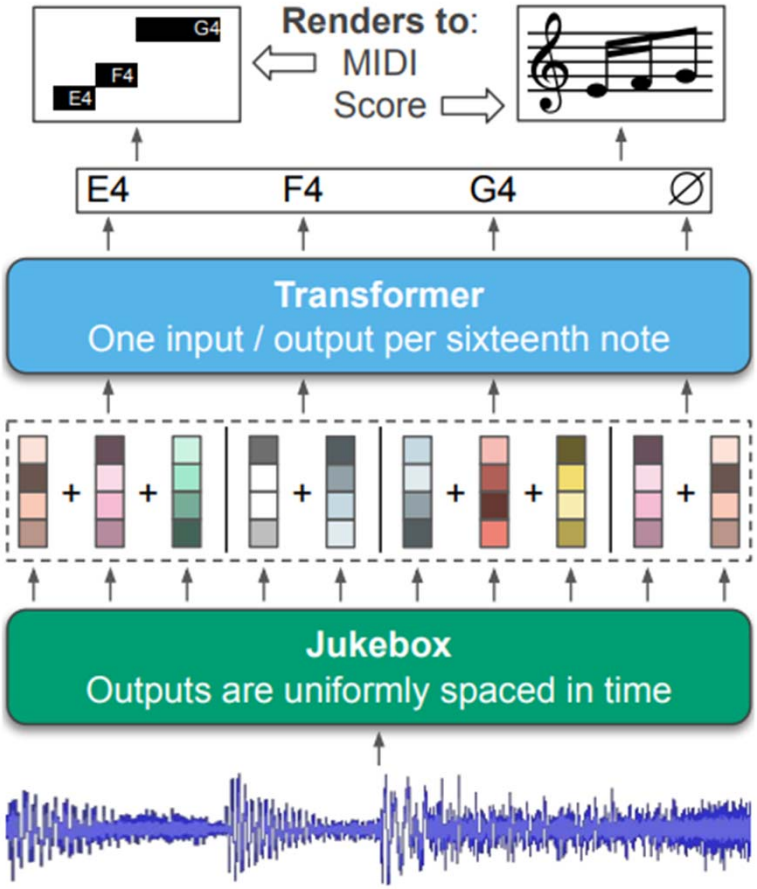


Lu et al., "Multitrack music transcription with a time-frequency perceiver," ICASSP 2023

Exemplar Model: Sheet Sage (for Lead Sheet Transcription)

https://github.com/chrisdonahue/sheetsage

- Predict both melody and chord
  - Use a large pre-trained model called **Jukebox** as model backbone and then do transfer learning
  - Use **Transformer** to learn the language model (LM) for melody and chords
  - Computationally heavy but pretty accurate
  - Lighter alternatives: BTC (https://github.com/jayg996/BTC-ISMIR19)



Donahue et al., "Melody transcription via generative pre-training," ISMIR 2022

Guitar Transcription: Need to be Invariant to Audio Effects

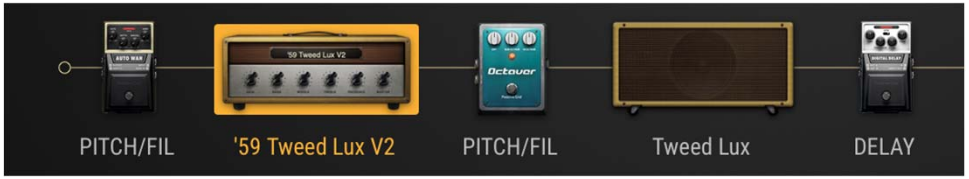
(Examples provided by Positive Grid®)

Dry (single coil EG)

In the Clouds

Overdriven Verb Icon

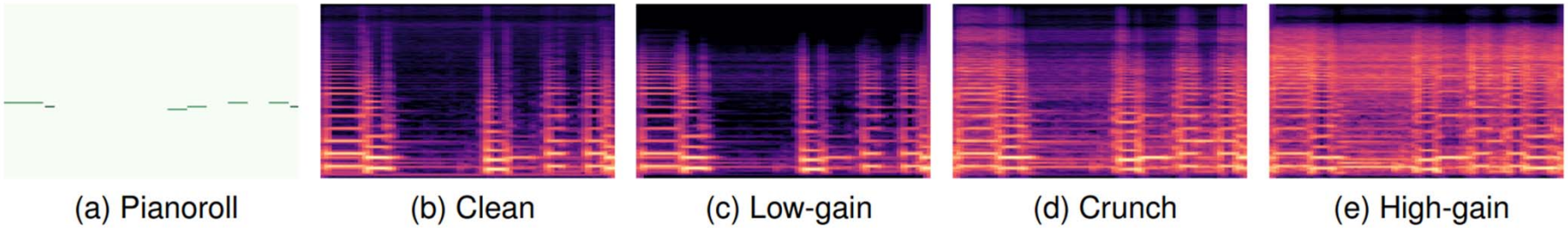
Lazy Down



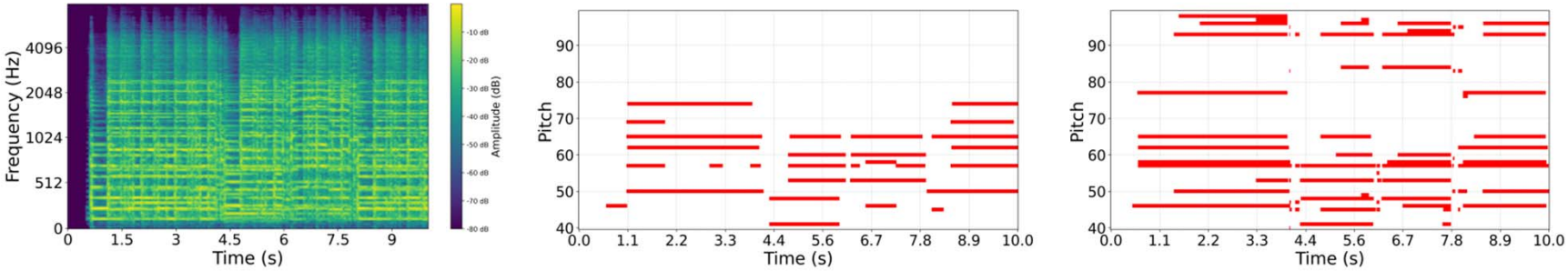
Chen et al., "Towards automatic transcription of polyphonic electric guitar music: A new dataset and a multi-loss transformer model," ICASSP 2022

Guitar Transcription: Need to be Invariant to Audio Effects

https://ss12f32v.github.io/Guitar-Transcription-with-Amplifier/



(a) Pianoroll (b) Clean (c) Low-gain (d) Crunch (e) High-gain



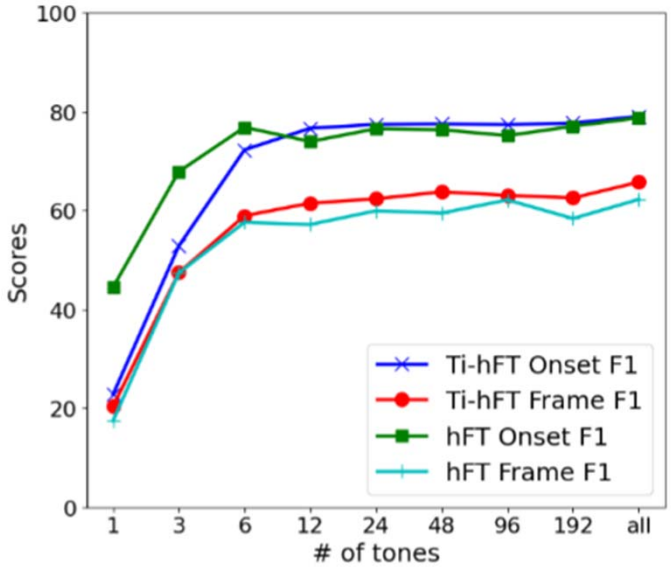
(a) log-mel spectrogram (b) TIT (c) hFT-Maestro-EGDB-PG

Chen et al., “Towards generalizability to tone and content variations in the transcription of amplifier rendered electric guitar audio,” arXiv 2025

Guitar Transcription: Need to be Invariant to Audio Effects

- Approach 1: make the audio input **tone-normalized**
  - However, the performance of existing tone removal models is not robust enough
- Approach 2: make the model **tone-informed**
  - Effective with large number of seen tones

Model	High-gain	
	Onset F1	Frame F1
Ti-hFT (proposed)	78.9	**64.5**
hFT-Maestro [11]	42.6	39.1
hFT-Maestro-E&G-PG [11]	51.3	45.2
MT3-Guitar [17]	46.8	7.8
MT3-Guitar&Piano [17]	70.0	10.0
MT3-All [17]	**86.1**	8.9
TabCNN [16]	n/a	37.7
TabCNNx4 [12]	n/a	40.9



(c) High-gain

Chen et al., "Towards generalizability to tone and content variations in the transcription of amplifier rendered electric guitar audio," arXiv 2025

ISMIR 2021 Tutorial: Programming MIR Baselines from Scratch - Pitch Tracking

https://github.com/rabitt/ismir-2021-tutorial-case-studies/tree/main/pitch_tracking

- Video online

  https://drive.google.com/file/d/18LNaKy2ymFjEWj19gHy25wgtFV4pdML0/view