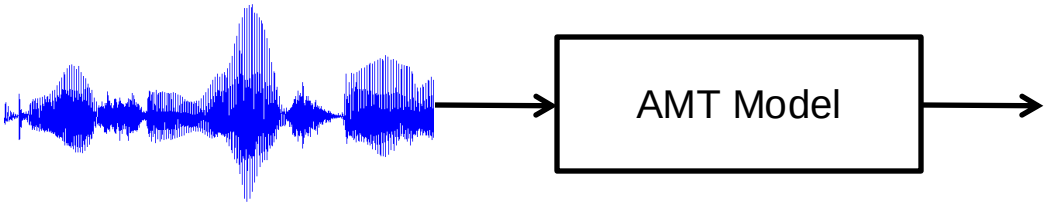
GCT634/AI613: Musical Applications of Machine Learning

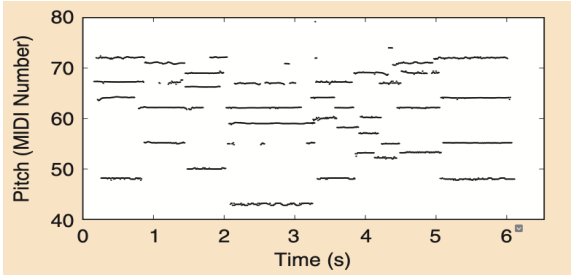
Automatic Music Transcription: Polyphonic

Juhan Nam

Multiple Pitch Estimation

- Polyphonic pitch estimation from multiple sound sources





Challenges

- Many sources are mixed and played simultaneously
  - They are likely to be harmonically related in music
  - Some sources can be masked by others
  - Content changes continuously by musical expressions (e.g. vibrato)

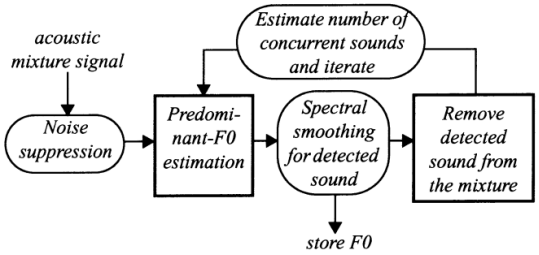
- Labeling is time-consuming and requires high expertise
  - Supervised learning is limited (piano transcription is a special case)
  - Sheet music can be used as “weak” labels with the score-to-audio alignment
  - Multi-track recording with monophonic pitch estimation

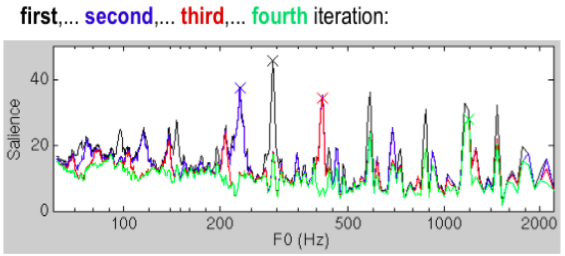
Methods

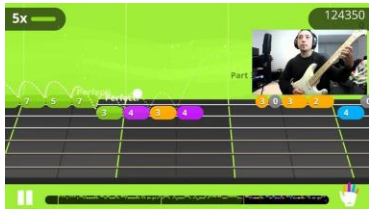
- Iterative F0 search: DSP
- Joint source estimation: NMF
- Classification-based approach: ML/DL

Iterative F0 search

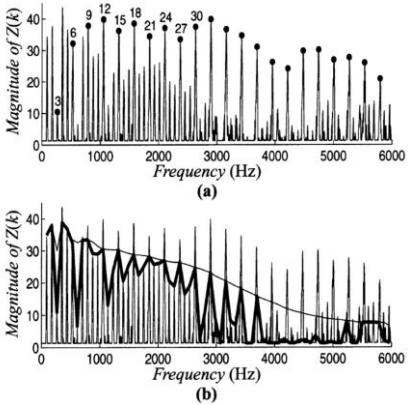
- Repeatedly finds predominant-F0 and removes its harmonic overtones
- Procedure
  1. Set the original to the residual
  2. Detect a predominant F0: based on the pitch templates
  3. Spectral smoothing on harmonics on the detected F0
  4. Cancel the smoothed harmonics from the residual
  5. Repeat the step 2 & 3 until the residual is sufficiently flat







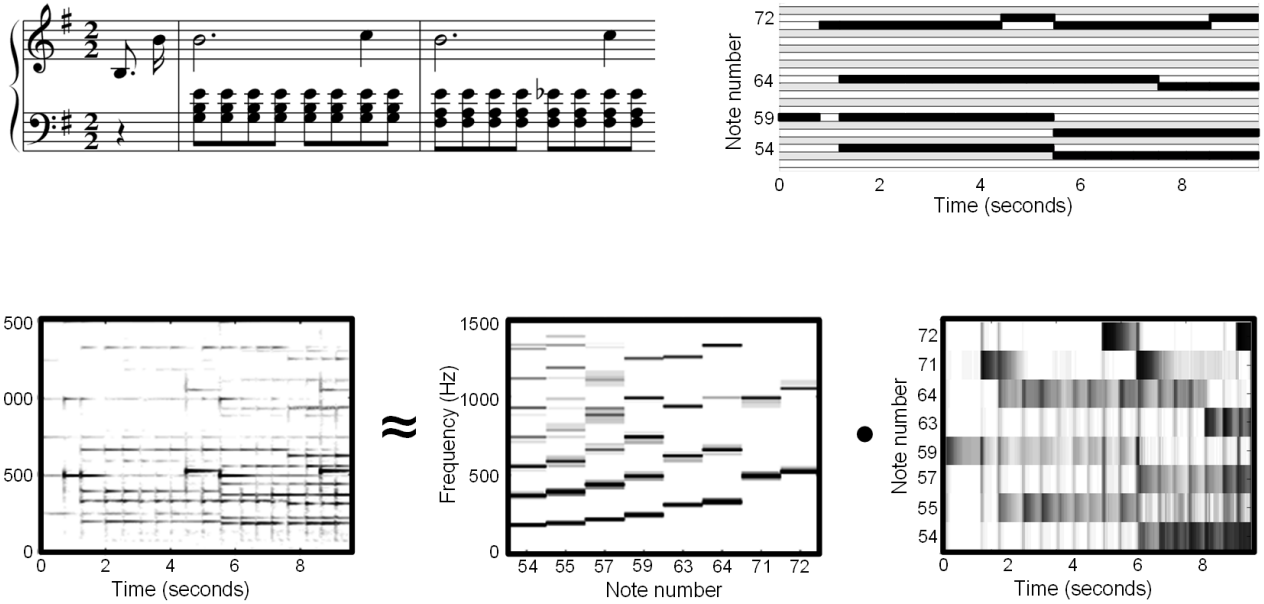
Yousician



Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness, Anssi Klapuri, IEEE TASLP, 2003

NMF-based Spectrogram Decomposition

- Spectrogram can be approximated with an additive sum of pitch templates and the corresponding temporal activations

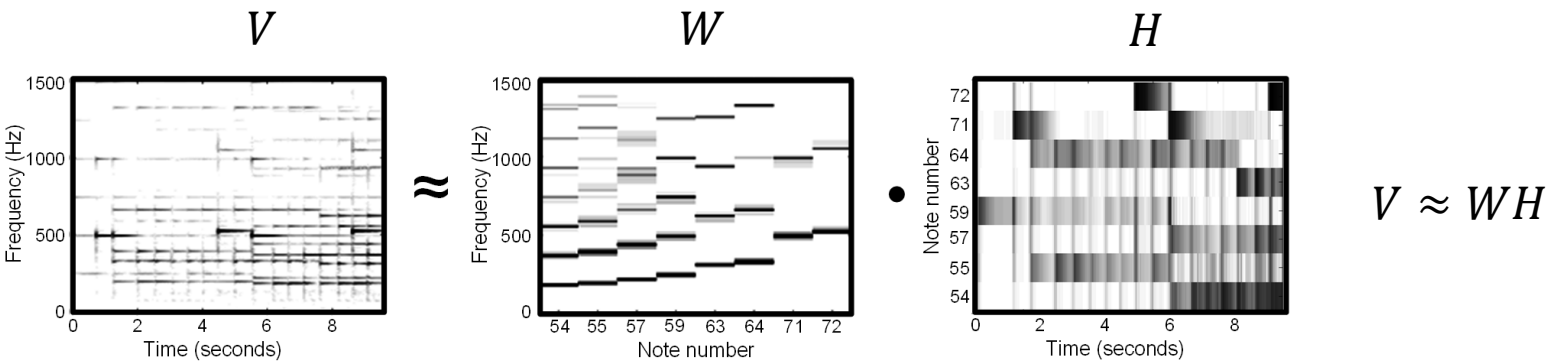


[The FMP book]

NMF-based Spectrogram Decomposition

- They can be regarded as a non-negative matrix factorization
  - All elements are non-negative

$$F \times T \approx F \times K \times K \times T$$



Non-Negative Matrix Factorization for Polyphonic Music Transcription, Paris Smaragdis, Judith Brown, WASPAA, 2003

Algorithm for NMF

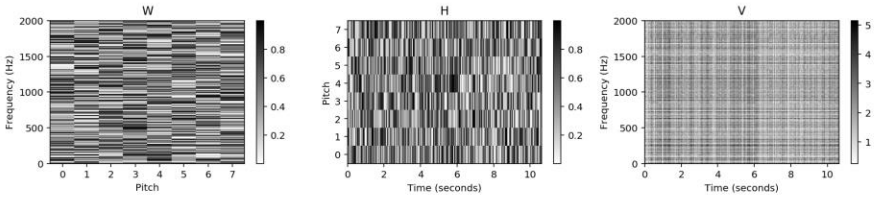
- Defined as an optimization problem: $\min_{W,H \geq 0} D(V||WH)$
  - Euclidean: $D(V||\hat{V}) = \sum_{i,j}(V_{ij} - \hat{V}_{ij})^2 \quad (\hat{V} \approx WH)$
  - Kullback-Leibler (KL) divergence: $D(V||\hat{V}) = \sum_{i,j}(V_{ij} log \frac{V_{ij}}{\hat{V}_{ij}} - \hat{V}_{ij} + V_{ij})$
- Multiplicative update rule
  1. Initialize W and H
  2. Repeat $H \leftarrow H .* \frac{W^T \frac{V}{WH}}{W^T 1} \qquad W \leftarrow W .* \frac{\frac{V}{WH} H^T}{1 H^T}$
  3. Until convergence
  4. Return W and H

Algorithms for Non-negative Matrix Factorization, Daniel Lee, Sebastian Seung, NIPS, 2000

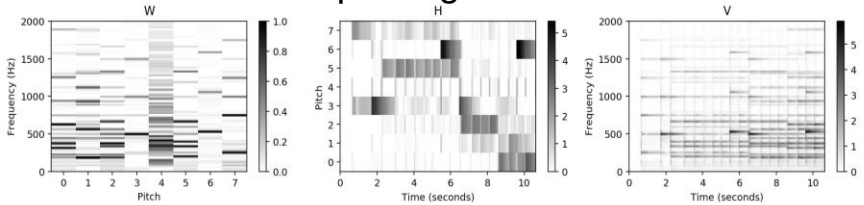
NMF for Polyphonic Pitch Estimation

- Initialize W with harmonic template

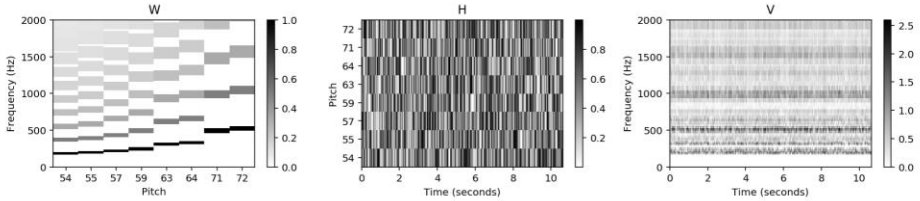
Random Initially W and H



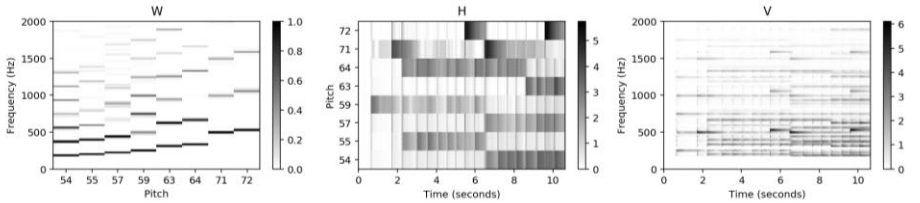
After updating W and H



Initialize W and H with harmonic templates



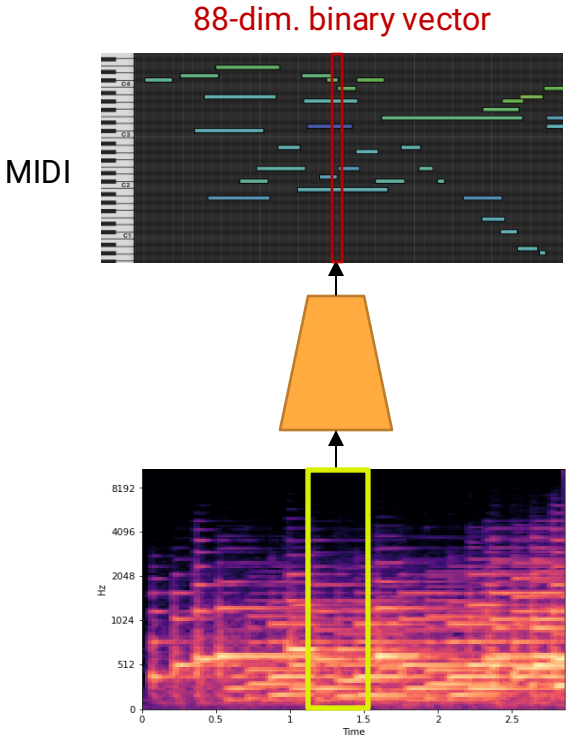
After updating W and H



- NMF examples
  - https://www.audiolabs-erlangen.de/resources/MIR/FMP/C8/C8.html (8.3)

Classification-based Approach

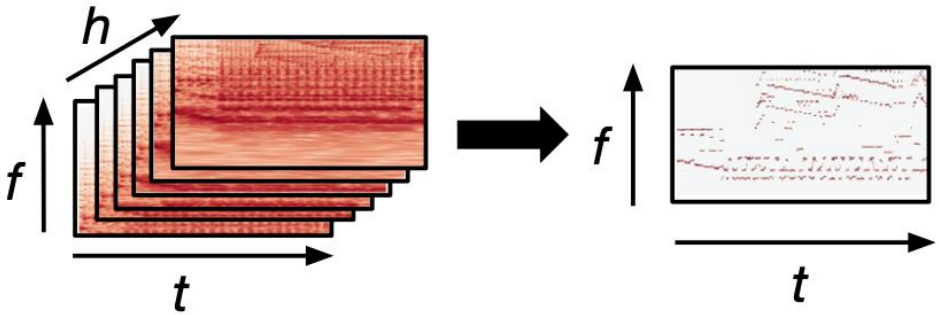
- Quantize the pitch output into discrete label vectors
- Multi-label classification
  - 88 binary state output (note on/off)
  - Use the sigmoid output
- No prior knowledge of musical acoustics



Classification-based Multi-Pitch Estimation

- Predict the pitch saliency from multi-track instruments
  - Frame-level pitch activations in the time and pitch space
- Input representation: harmonic constant-Q transform (HCQT)
  - CQT with 60 bins per octave
  - Multiple CQTs with harmonic relations (0.5, 1, 2, 3, 4, 5)
  - Filters learn the relative weights of harmonics
  - 3D input (time x frequency x harmonics): similar to color images (RGB)

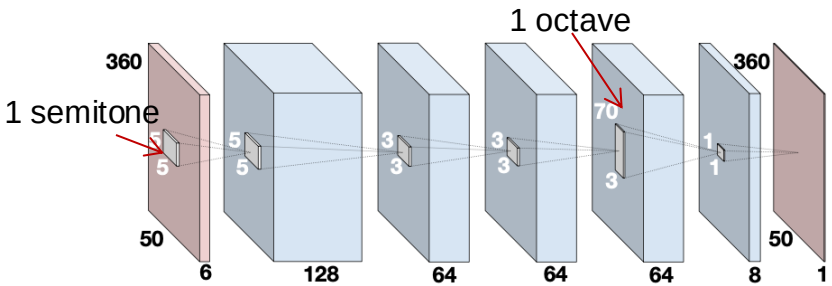
HCQT is similar to the idea of **harmonic product sum** but they stack them as different channels

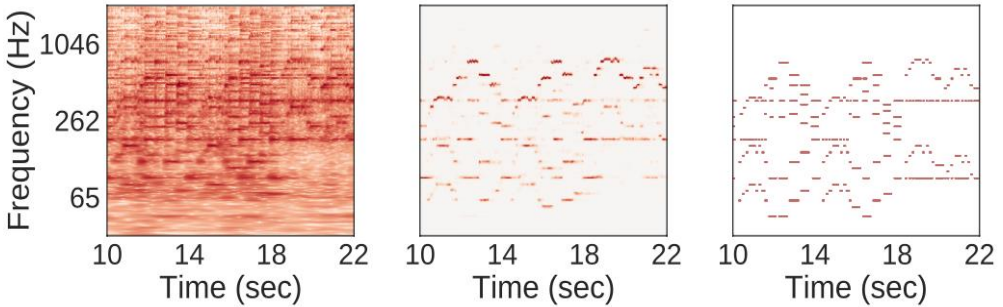


Deep salience representations for F0 estimation in polyphonic music Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, Juan P. Bello, ISMIR, 2017

Classification-based Multi-Pitch Estimation

- 2D CNN
  - 5x5 filter: 1 semitone, 70 x 3: one octave
  - The output layer has a sigmoid output
    - The loss is cross-entropy between the sigmoid output and the ground truth
    - They used the Gaussian blurring (smoothing) function on the ground truth
  - ReLU, batch norm, Adam optimizer
  - The input an output have the same dimensionality: no pooling layers

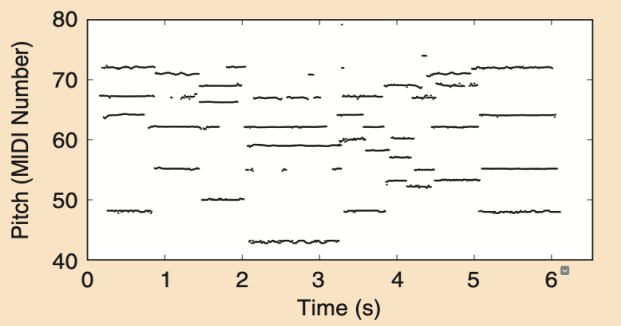


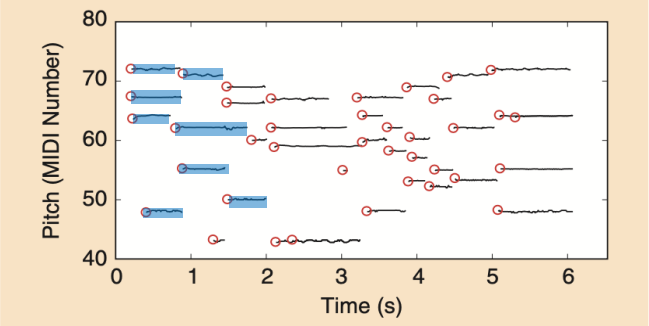


Deep salience representations for F0 estimation in polyphonic music Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, Juan P. Bello, ISMIR, 2017

Note-Level Transcription

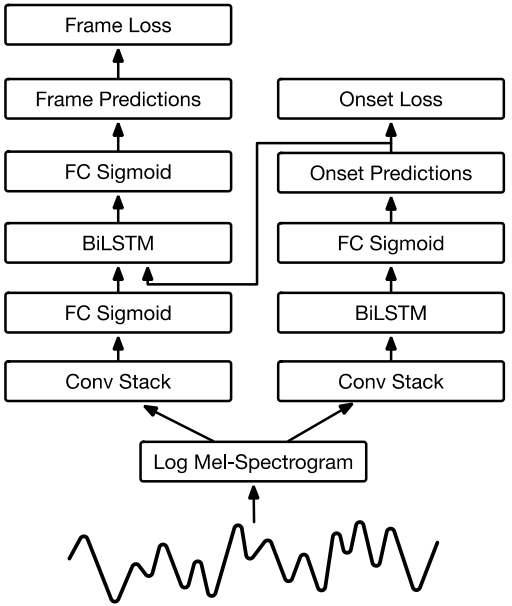
- Convert continuous pitch streams into note events
  - Use the frame-level pitch estimation
  - Explicit onset detectors can be added but they are very hard
    - The classification-based approach is common nowadays
  - Note modeling algorithms to prune, merge, and divide frame-level predictions
    - Rule-based approach: thresholding, median filtering
    - Statistical approach: HMM





Onsets and Frames

- Joint learning of onset detection and pitch estimation for polyphonic piano transcription
  - Two CRNN branches
    - Onset network: detect the onset of multiple notes (percussive tone)
    - Frame network: detect on/off states of multiple notes (harmonic tone)
  - A connection from the onset prediction in the onset network to the input of RNN in the frame network
    - Temporal causality

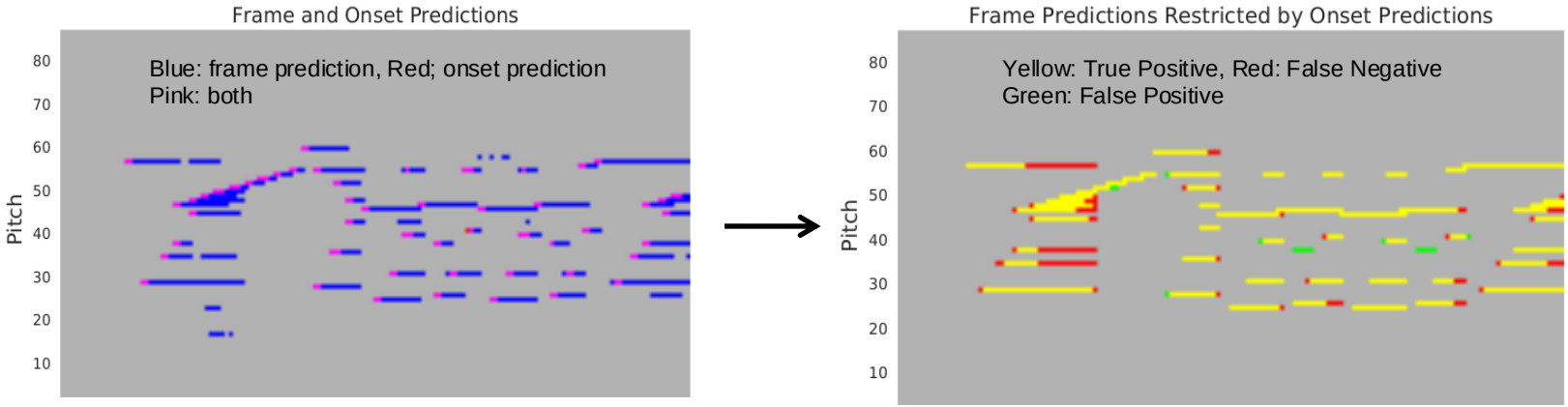


Website: https://magenta.tensorflow.org/onsets-frames

Onsets and Frames: Dual-Objective Piano Transcription, Curtis Hawthorne, et al, ISMIR, 2018

Onsets and Frames

- A simple rule is used to integrate the output of the two networks
  - Frame predictions without onset is discarded



Onsets and Frames: Dual-Objective Piano Transcription, Curtis Hawthorne, et al, ISMIR, 2018

Onsets and Frames

- Significantly outperform the previous state-of-the-arts
  - High jump in the note-level accuracy
  - The key idea is detecting “onset” state separately
  - The following studies investigated more note states: onset, sustain, and offset, and even detection of the sustain pedal

	Frame			Note			Note w/ offset			Note w/ offset & velocity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Sigtia et al., 2016 [18]	71.99	**73.32**	72.22	44.97	49.55	46.58	17.64	19.71	18.38	—	—	—
Kelz et al., 2016 [13]	81.18	65.07	71.60	44.27	61.29	50.94	20.13	27.80	23.14	—	—	—
Melodyne (decay mode)	71.85	50.39	58.57	62.08	48.53	54.02	21.09	16.56	18.40	10.43	8.15	9.08
Onsets and Frames	**88.53**	70.89	**78.30**	**84.24**	**80.67**	**82.29**	**51.32**	**49.31**	**50.22**	**35.52**	**30.80**	**35.39**

Table 1. Precision, Recall, and F1 Results on MAPS configuration 2 test dataset (ENSTDkCl and ENSTDkAm full-length .wav files). Note-based scores calculated by the *mir_eval* library, frame-based scores as defined in [2]. Final metric is the mean of scores calculated per piece. MIDI files used to calculate these scores are available at `https://goo.gl/magenta/onsets-frames-examples`.

	F1		
	Frame	Note	Note with offset
Onset and Frames	**78.30**	**82.29**	**50.22**
(a) Frame-only LSTM	76.12	62.71	27.89
(b) No Onset Inference	**78.37**	67.44	34.15
(c) Onset forward LSTM	75.98	80.77	46.36
(d) Frame forward LSTM	76.30	82.27	49.50
(e) No Onset LSTM	75.90	80.99	46.14
(f) Pretrain Onsets	75.56	81.95	48.02
(g) No Weighted Loss	75.54	80.07	48.55
(h) Shared conv	76.85	81.64	43.61
(i) Disconnected Detectors	73.91	**82.67**	44.83
(j) CQT Input	73.07	76.38	41.14
(k) No LSTM, shared conv	67.60	75.34	37.03

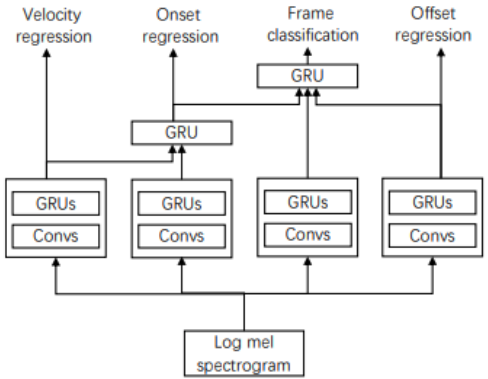
Table 2. Ablation Study Results.

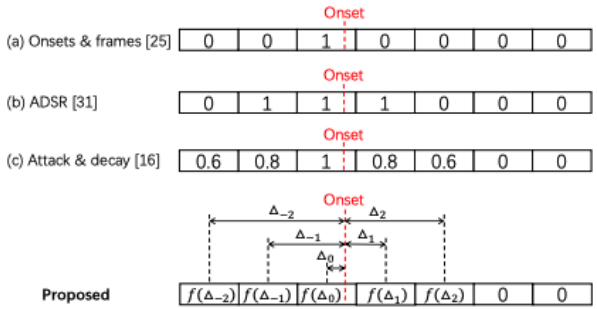
Onsets and Frames: Dual-Objective Piano Transcription, Curtis Hawthorne, et al, ISMIR, 2018

Regression Onset Model

- High-resolution piano transcription using the regression loss
  - The ground truth of onset and offset are smoothed using a triangular shape
    - But, use the binary cross entropy
  - Achieve more precise onset and offset prediction than the onsets and frames model (hop size: 32ms)

Website: https://github.com/bytedance/piano_transcription





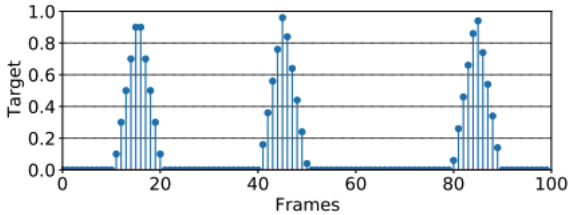
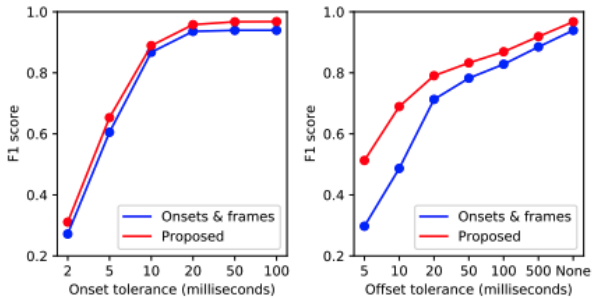


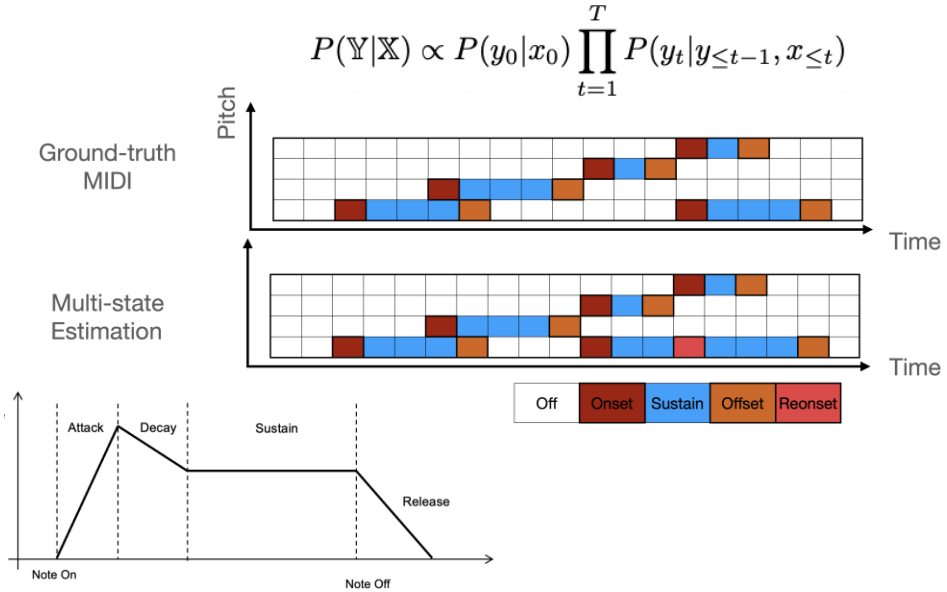
Fig. 2. High-resolution training targets of three notes with a same pitch.

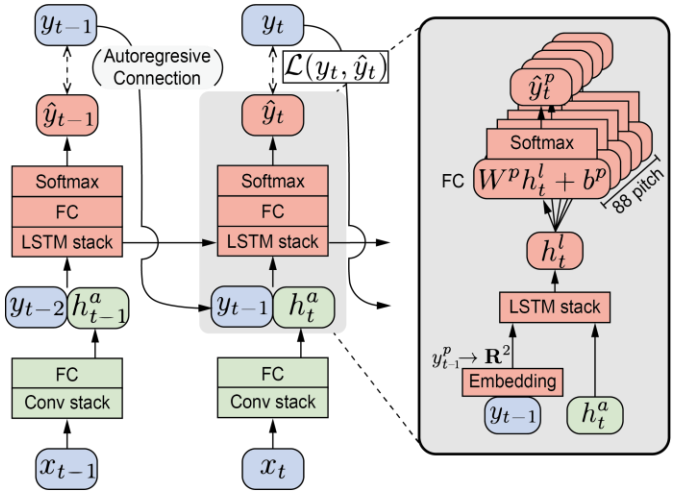


High-resolution Piano Transcription with Pedals by Regressing Onset and Offset Times, Qiuqiang Kong, et al., IEEE TASLP, 2021

Autoregressive Multi-State Note Model

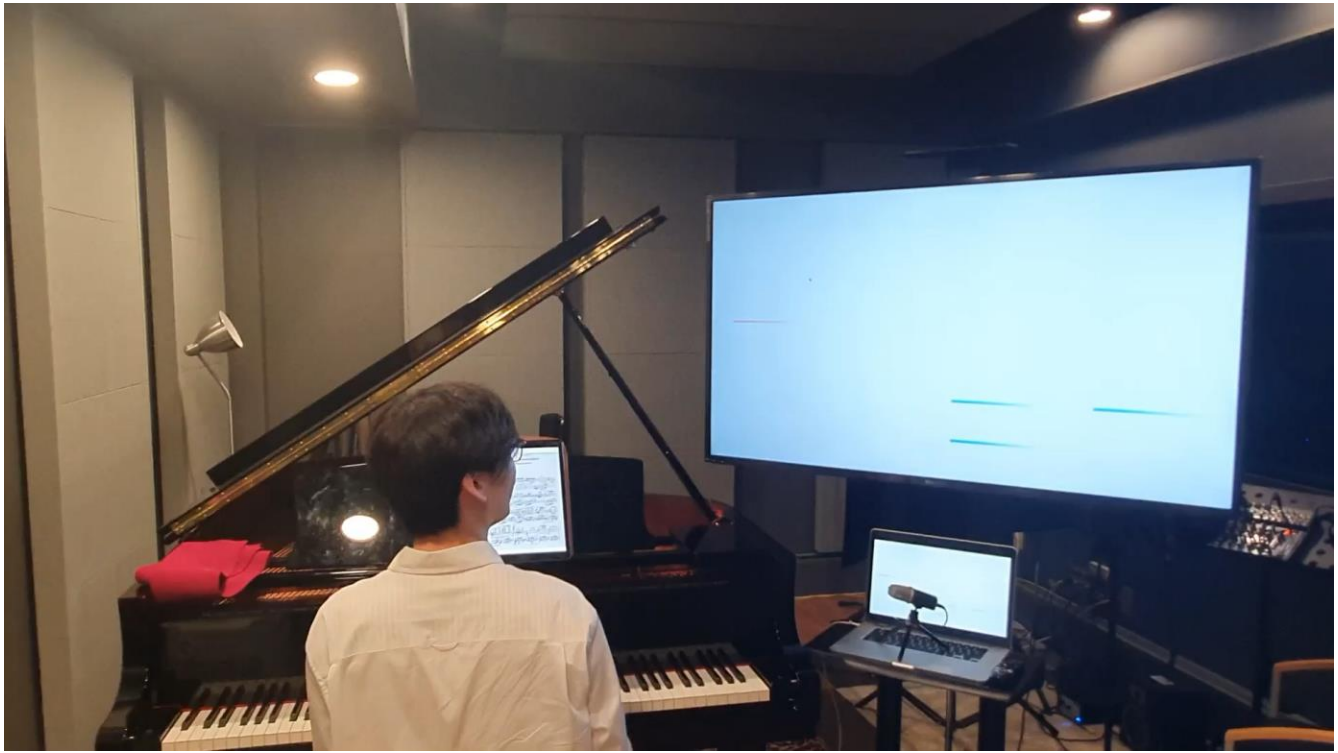
- Use a single CRNN with the softmax output that predicts multiple note states at once (off, onset, sustain, offset, and re-onset)
  - Autoregressive unidirectional RNN → real-time inference





Polyphonic Piano Transcription Using Autoregressive Multi-State Note Model, Taegyun Kwon, Dasaem Jeong, and Juhan Nam, ISMIR, 2020

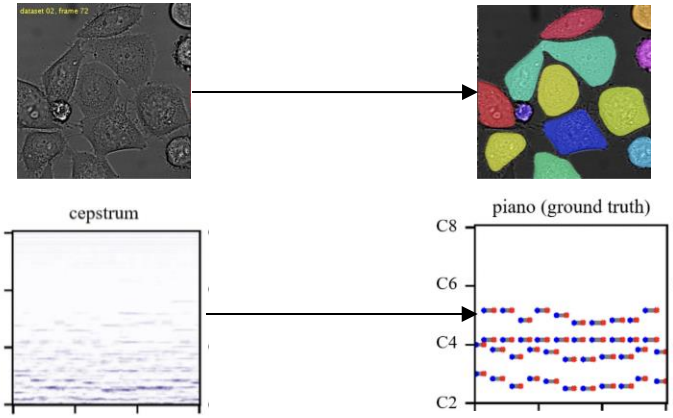
Demo: Real-Time Polyphonic Piano Transcription

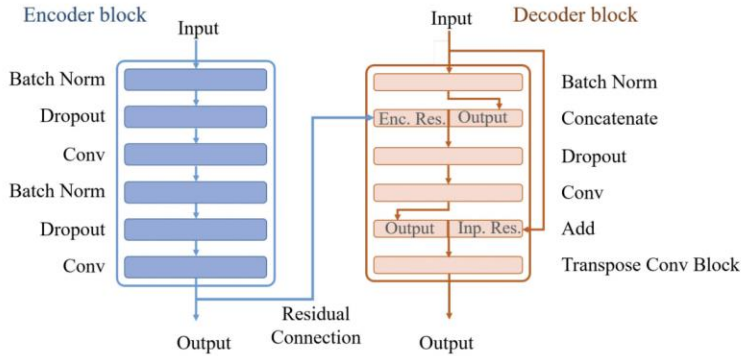


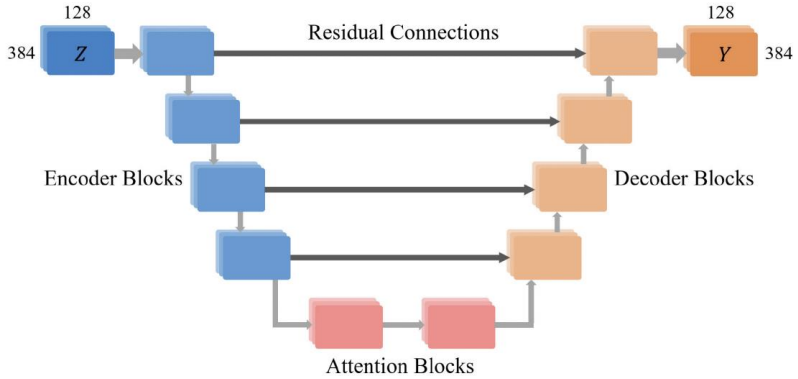
Music and Audio Computing Lab, KAIST (2020)

U-Net based Multi-Instrument AMT

- CNN-based Encoder-Decoder
  - Proposed for image segmentation
  - Use it for “note segmentation”
  - Self-attention for instrument detection



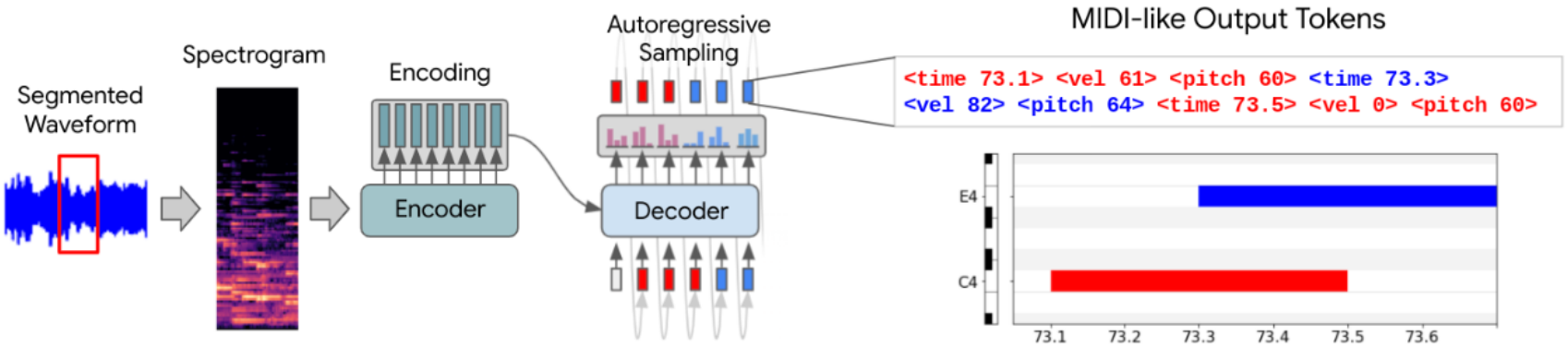




Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation, Yu-Te Wu, Berlin Chen, and Li Su, IEEE TASLP, 2020

Seq-to-Seq Model

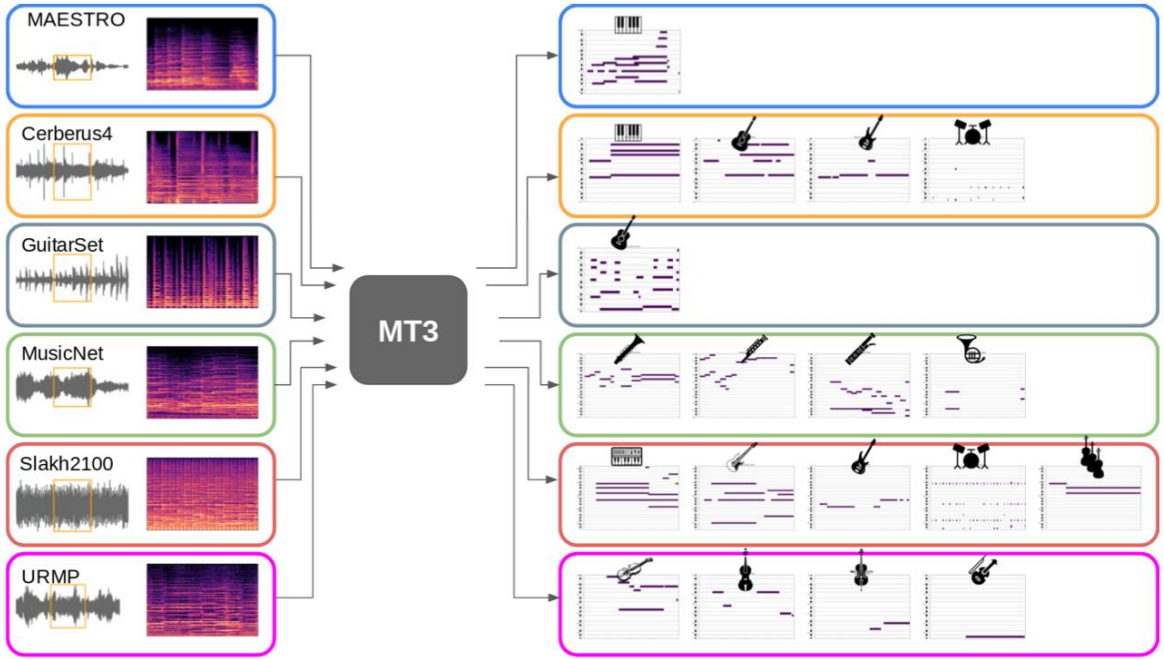
- A generic encoder-decoder Transformer with standard decoding methods
  - Represents the MIDI output with text-based token sequences



Sequence-to-Sequence Piano Transcription with Transformers, Curtis Hawthorne, Ian Simon, Rigel Swavely, Ethan Manilow, Jesse Engel, ISMIR, 2021

MT3

- The same seq-to-seq model that supports multi-task AMT

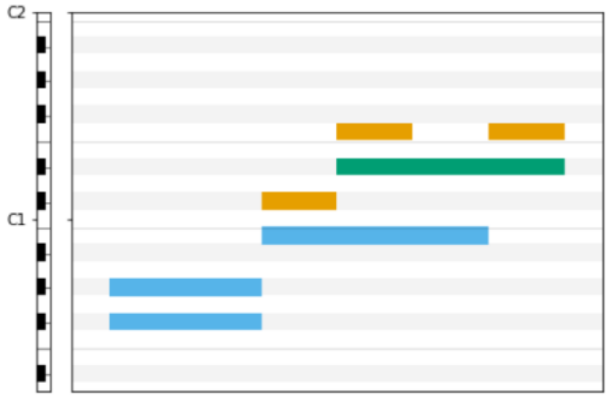


MT3: Multi-Task Multitrack Music Transcription, Josh Gardner, Ian Simon, Ethan Manilow†, Curtis Hawthorne, Jesse Engel, ICML, 2022

MT3

- Add the **“program change” token** to the output to change instruments
  - This allows the model to handle an arbitrary number of instruments

Piano Roll



Tokenization

Detokenization

MIDI-Like Target/Output Tokens

```
<time 0.27> <prog 26> <on>  <note 22>
                            <note 24>
<time 0.63>           <off> <note 22>
                            <note 24>
            <prog 0>  <on>  <note 29>
            <prog 26>       <note 27>
<time 1.55> <prog 32>       <note 31>
<time 1.60> <prog 0>  <off> <note 29>
                      <on>  <note 33>
<time 1.75>           <off> <note 33>
<time 1.90> <prog 26>       <note 27>
<time 2.03> <prog 32>       <note 31>
```

MT3: Multi-Task Multitrack Music Transcription, Josh Gardner, Ian Simon, Ethan Manilow†, Curtis Hawthorne, Jesse Engel, ICML, 2022

MT3

Model	MAESTRO	Cerberus4	GuitarSet	MusicNet	Slakh2100	URMP
			Frame F1			
Hawthorne et al. (2021)	0.66	–	–	–	–	–
Manilow et al. (2020)	–	0.63	0.54	–	–	–
Cheuk et al. (2021)	–	–	–	0.48	–	–
Melodyne	0.41	0.39	0.62	0.13	0.47	0.30
MT3 (single dataset)	**0.88**	0.85	0.82	0.60	0.78	0.49
MT3 (mixture)	0.86	**0.87**	**0.89**	**0.68**	**0.79**	**0.83**
			Onset F1			
Hawthorne et al. (2021)	0.96	–	–	–	–	–
Manilow et al. (2020)	–	0.67	0.16	–	–	–
Cheuk et al. (2021)	–	–	–	0.29	–	–
Melodyne	0.52	0.24	0.28	0.04	0.30	0.09
MT3 (single dataset)	**0.96**	0.89	0.83	0.39	0.76	0.40
MT3 (mixture)	0.95	**0.92**	**0.90**	**0.50**	**0.76**	**0.77**
			Onset+Offset F1			
Hawthorne et al. (2021)	0.84	–	–	–	–	–
Manilow et al. (2020)	–	0.37	0.08	–	–	–
Cheuk et al. (2021)	–	–	–	0.11	–	–
Melodyne	0.06	0.07	0.13	0.01	0.10	0.04
MT3 (single dataset)	**0.84**	0.76	0.65	0.21	0.57	0.16
MT3 (mixture)	0.80	**0.80**	**0.78**	**0.33**	**0.57**	**0.58**
Mixture (Δ%)	-5.3	+5.2	+19.5	+54.0	+0.1	+263

MT3: Multi-Task Multitrack Music Transcription, Josh Gardner, Ian Simon, Ethan Manilow†, Curtis Hawthorne, Jesse Engel, ICML, 2022

Datasets

- Piano
  - MAESTRO: large-scale real performance
    - https://magenta.tensorflow.org/datasets/maestro
  - MAPS: synthesized piano
    - https://adasp.telecom-paris.fr/resources/2010-07-08-maps-database/
  - Saarland Music Data (SMD): real performance
    - https://resources.mpi-inf.mpg.de/SMD/SMD_MIDI-Audio-Piano-Music.html

Datasets

- Multi instrument

Dataset	Hrs. Audio	Num. Songs	Num. Instr.	Instr. Per Song	Align	Low-Resource	Synthetic	Drums
Slakh2100	969	1405	35	4–48	Good		✓	✓
Cerberus4	543	1327	4	4	Good		✓	✓
MAESTROv3	199	1276	1	1	Good			
MusicNet	34	330	11	1–8	Poor	✓		
GuitarSet	3	360	1	1	Good	✓		
URMP	1	44	14	2–5	Fair	✓		

Dataset Name	Style / Instrumentation	Pitch annotation strategy	Mix Tracks	Works[1]	Versions[2]	hh:mm
MusicNet [18]	Chamber music (piano, strings, winds)	Aligned scores	330	306	1 up to 3	34:08
Schubert Winterreise [17]	Chamber music (piano, solo voice)	Aligned scores	216	24	9	10:50
TRIOS [14]	Chamber music (piano, strings, winds)	Multi-track	5	5	1	0:03
Bach10 [13]	Chamber music (violin, winds)	Multi-track	10	10	1	0:06
PHENICX-Anechoic [15]	Symphonic (orchestra)	Multi-track	4	4	1	0:10
Choral Singing Dataset [16]	A cappella (choir)	MIDI-guided performance	3	3	1	0:07

MT3: Multi-Task Multitrack Music Transcription, Josh Gardner, Ian Simon, Ethan Manilow†, Curtis Hawthorne, Jesse Engel, ICML, 2022

Deep-Learning Architectures for Multi-Pitch Estimation: Towards Reliable Evaluation, Christof Weiß, Geoffroy Peeters, Arxiv, 2022