Oregon State University

Transcription of Audio to MIDI using Deep Learning

Patrick Donnelly & Victoria Ebert

soundbendor lab

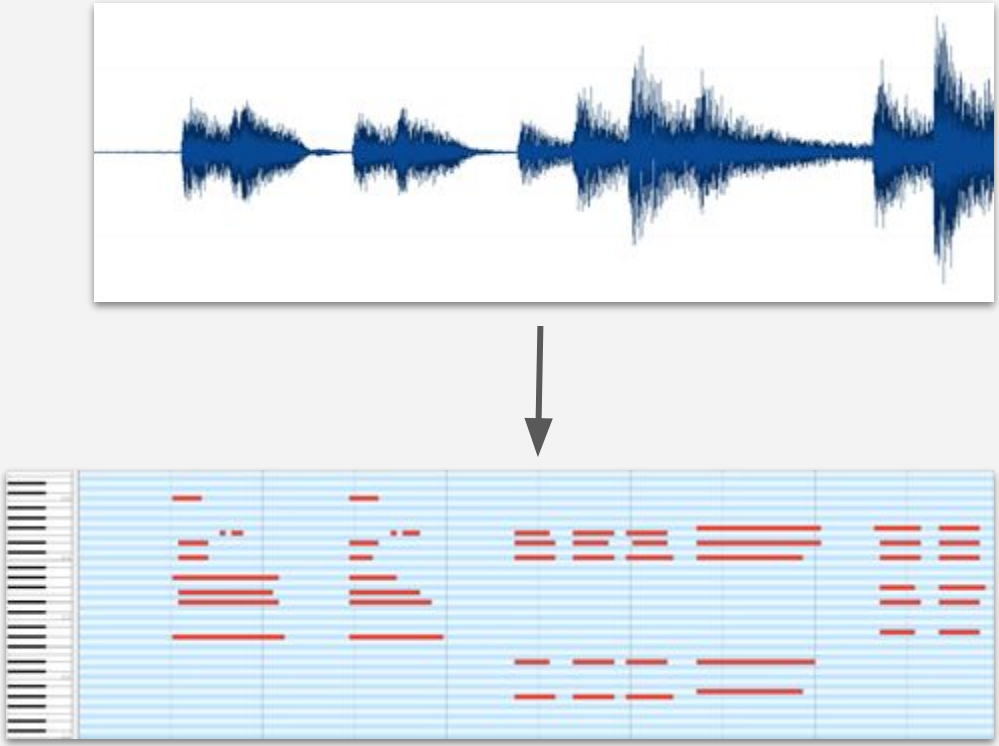
College of Engineering | Computer Science

Oregon State University

INTRODUCTION

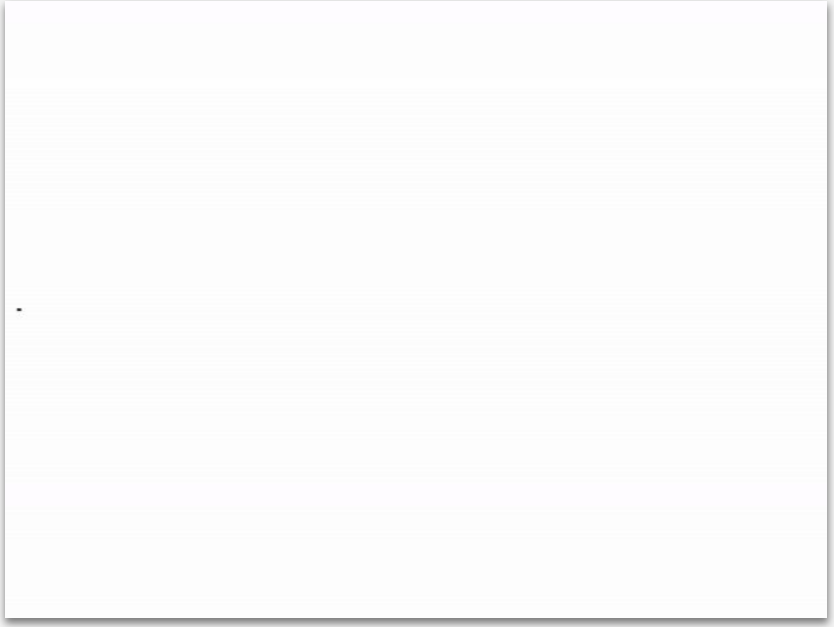
Music Transcription

Oregon State University



Problem Statement

- Automatic music transcription
  - audio ⇒ MIDI
  - deep-learning
  - raw audio as input
  - no feature extraction



Related Work

Oregon State University

Music Information Retrieval

- Generation
- Classification
- Alignment
- **Transcription**

Automatic Music Transcription

- neural network per note[1]
- support vector machine[2]
- CNN constant Q transform[3]
- frame level classification[4]
- onset detection[2-15]
- Spectrograms[6-7], [15-16]

Oregon State University

DATASET

disklavier®

THE ULTIMATE PIANO

- Enjoy on-demand performances by your favorite artists with the easy-to-use *Enspire Controller* App.
- Experience lifelike piano performances complete with countless other instruments - even vocals!
- Make your next next party a hit with built-in “fun for beginners” *SmartKey™* songs!
- Listen to over 30 channels of *original artist* music (including seasonal channels) with *Disklavier Radio™*.
- Host LIVE, in-home concerts by Elton John, Sarah McLachlan, The Piano Guys & more with *Disklavier TV*!
- Play with headphones or connect to your home audio system and enjoy music throughout your entire house!
- Unwind while your piano automatically updates and calibrates itself! All you have to do is have it tuned!



MAESTRO Dataset

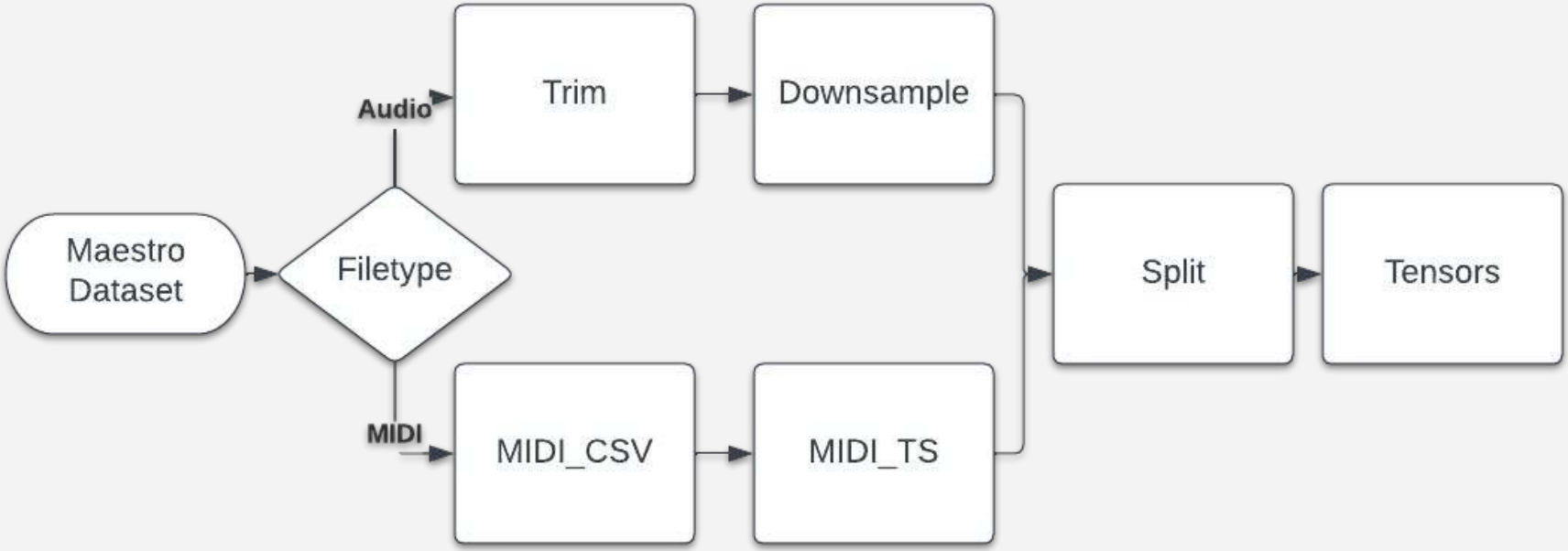
Oregon State University

- MIDI and Audio Edited for Synchronous TRacks and Organization
  - 10 years of International Piano E-Competition
  - released 2018
  - *concert hall*
  - *n* = 1276 recordings

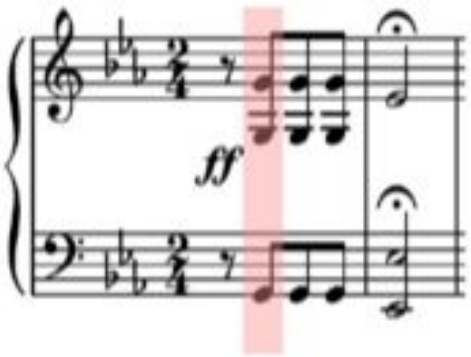
Split	**Performances** (count)	**Duration** (hours)	**Size** (GB)	**Notes** (millions)
Train	962	159.2	96.3	5.66
Validation	137	19.4	11.8	0.64
Test	177	20.0	12.1	0.74
Total	1276	198.7	120.2	7.04

Preprocessing

Oregon State University

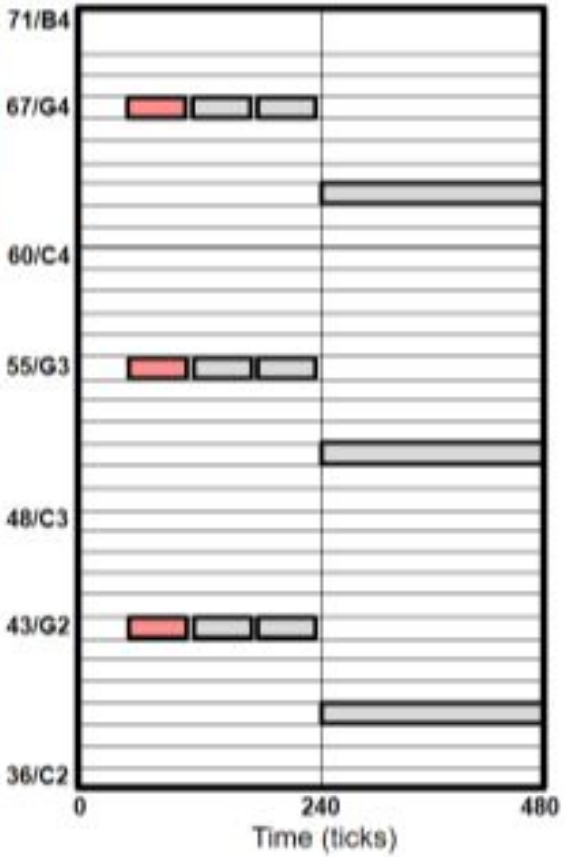


MIDI_CSV



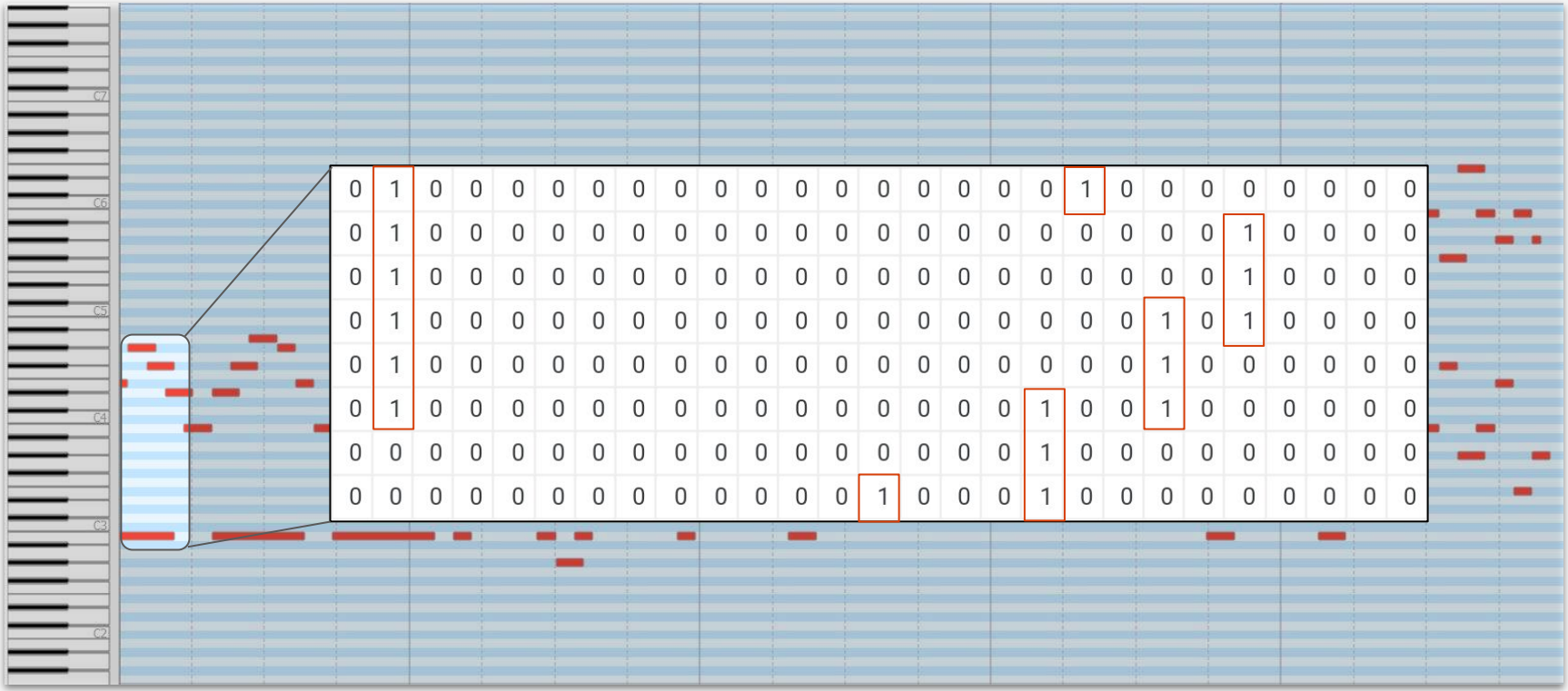
[17]

Time (Ticks)	Message	Channel	Note Number	Velocity
60	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	63	100
0	NOTE ON	2	51	100
0	NOTE ON	2	39	100
240	NOTE OFF	1	63	0
0	NOTE OFF	2	51	0
0	NOTE OFF	2	39	0



MIDI_TS

Oregon State University



Audio Processing

Oregon State University

- Downsample to 16 kHz
  - highest note is 4186 Hz
    - 3rd overtone is 16744 Hz
  - reduce input but keep auditory fidelity
- Split audio
  - 10 second chunks
  - convert to tensors

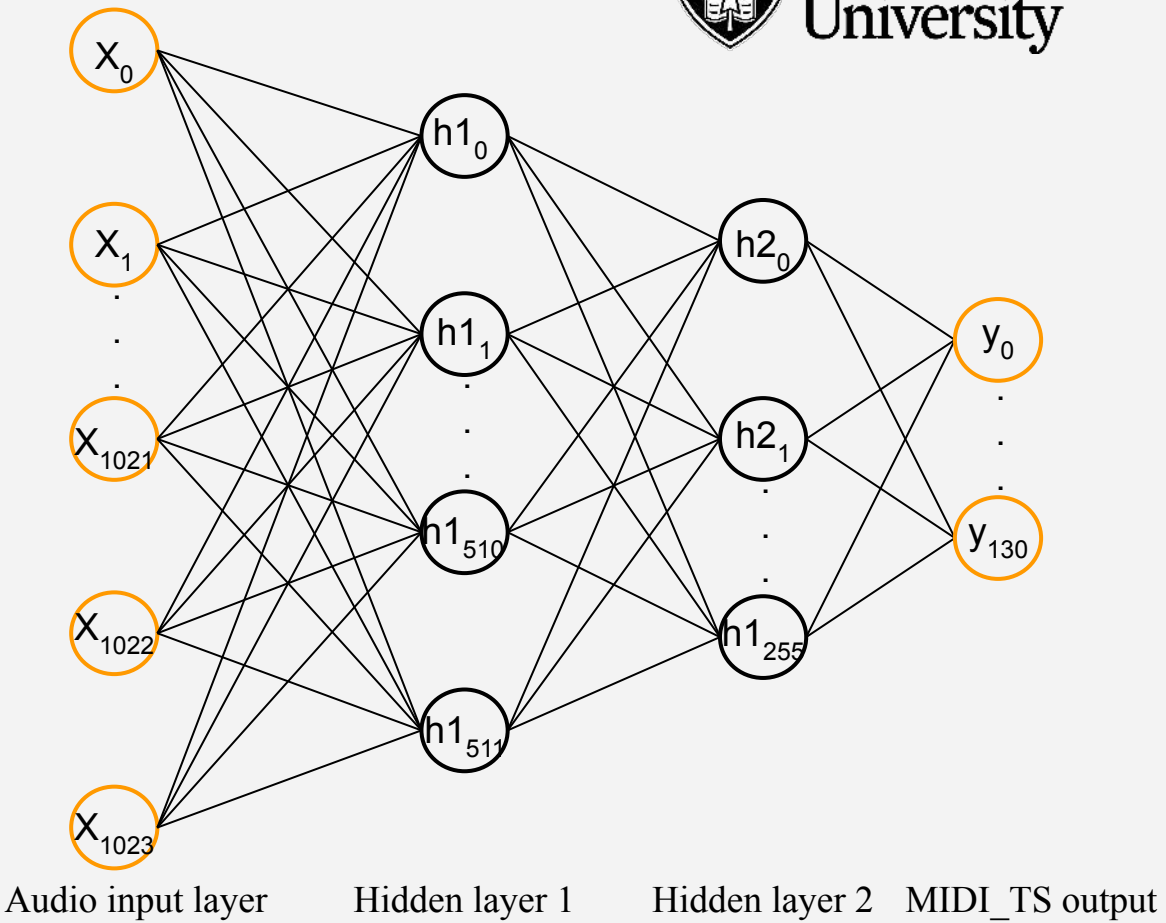


Oregon State University

EXPERIMENTAL DESIGN

Model Design

- BLSTM
  - input: **1024**
  - hidden: **512**
  - hidden: **256**
  - output: **131**
- dropout of 0.2
- loss using MSE



Post-Processing

- remove quiet noise
- smooth velocity for MIDI playback
- convert to MIDI_CSV then to MIDI

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0.064	0.000846	0	0	0	0.000249	0	0	0	0	0.000125	0	4.74E-05	2.57E-05	0.001037	3.25E-05	0.000678	0	0.006414
0.128	0.000584	6.28E-05	6.68E-05	2.59E-06	0	0	0	0	0	0.000208	0	0	1.39E-05	0	0.000224	0	0.000702	0.002077
0.192	0.000517	0.000226	6.56E-05	4.68E-05	0	0	1.20E-05	0	0	0.000127	8.42E-06	0	4.29E-06	0	0.000168	0	0.00111	0
0.256	0.000508	0.000233	3.67E-05	1.57E-05	0	0	2.92E-06	0.001088	0	7.35E-05	1.26E-06	0	1.33E-06	0	0.00012	0	0.000855	0
0.32	0.00051	0.000208	2.02E-05	0	0	0	0	0.002123	0	4.56E-05	0	0	1.30E-06	0	9.48E-05	0	0.000585	0
0.384	0.000522	0.00019	1.27E-05	0	0	0	0	0.002744	0	3.10E-05	0	0	2.44E-06	0	8.34E-05	0	0.00038	0
0.448	0.000571	0.000192	9.54E-06	0	0	0	0	0.003271	0	2.77E-05	0	0	4.41E-06	0	7.90E-05	0	0.000224	0
0.512	0.000657	0.000225	9.14E-06	0	0	0	0	0.003936	0	3.95E-05	0	0	7.34E-06	0	8.30E-05	0	9.18E-05	0
0.576	0.000673	0.000228	8.32E-06	0	0	0	0	0.0046	0	4.99E-05	0	0	8.16E-06	0	7.45E-05	0	0	0
0.64	0.00057	0.000204	1.05E-05	0	0	0	0	0.004693	0	4.29E-05	0	0	6.91E-06	1.36E-05	5.54E-05	0	0	0
0.704	0.000397	0.000224	1.45E-05	0	0	0	0	0.005304	0	4.22E-05	0	0	8.40E-06	1.79E-05	5.70E-05	0	0	0
0.768	0.0001	0.000235	1.36E-05	0	0	0	0	0.006527	0	3.69E-05	0	0	1.01E-05	8.72E-06	6.07E-05	0	0	0
0.832	0	0.000244	2.21E-05	0	0	9.83E-06	0	0.00876	0	2.20E-05	0	0	1.39E-05	0	5.92E-05	0	0	0.000381
0.896	0	0.000249	2.96E-05	0	0	2.65E-05	0	0.011599	0	0	0	0	2.08E-05	0	8.09E-05	0	0	0.000805
0.96	0	0.000286	3.03E-05	0	0	2.42E-05	0	0.012547	0	0	0	0	3.64E-05	0	0.000208	0	0.000228	0.001032
1.024	0	0.000225	0	0	0	0	0	0.004501	0	3.24E-05	0	0	4.09E-05	4.22E-06	0.000301	0.000114	0.000122	0.000801
1.088	0.000628	0.000137	0	0	0	0	0	0.002286	0	1.99E-05	0	0	1.92E-05	0.000299	0.000217	0.000208	0	0
1.152	0.000925	3.08E-05	0	0	0	0	0	0.001662	0	0	0	0	1.54E-05	0.00034	0.000142	0.000174	0	0
1.216	0.001033	7.18E-06	9.94E-06	0	0	0	0	0.00188	3.52E-06	0	0	0	1.33E-05	0.000329	0.00011	0.000156	0	7.02E-05

Evaluation Metrics:

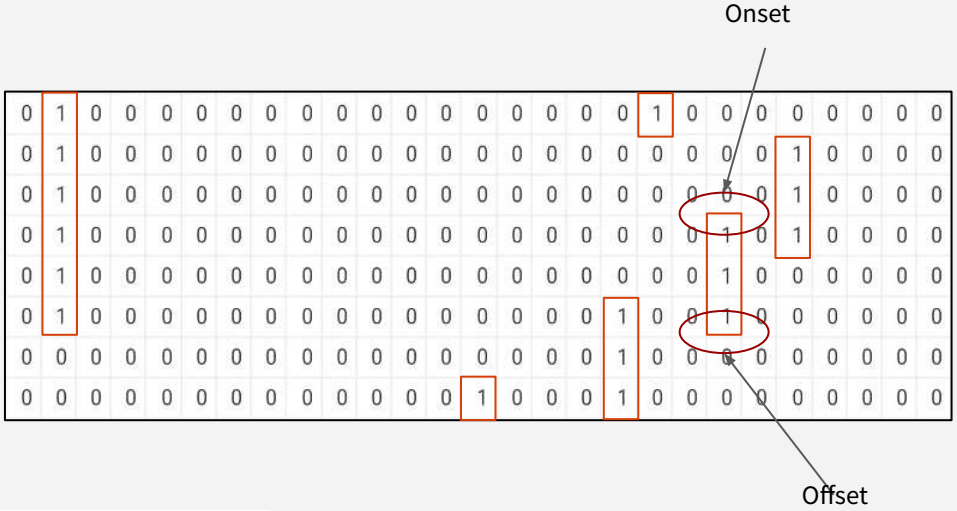
Oregon State University

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$f - measure = 2 * \frac{precision * recall}{precision + recall}$$

$$overlap\ ratio = \frac{(min(true_offset, predicted_offset) - max(true_onset, predicted_onset))}{(max(true_offset, predicted_offset) - min(true_onset, predicted_onset)))}$$



`mir_eval`

Oregon State University

RESULTS

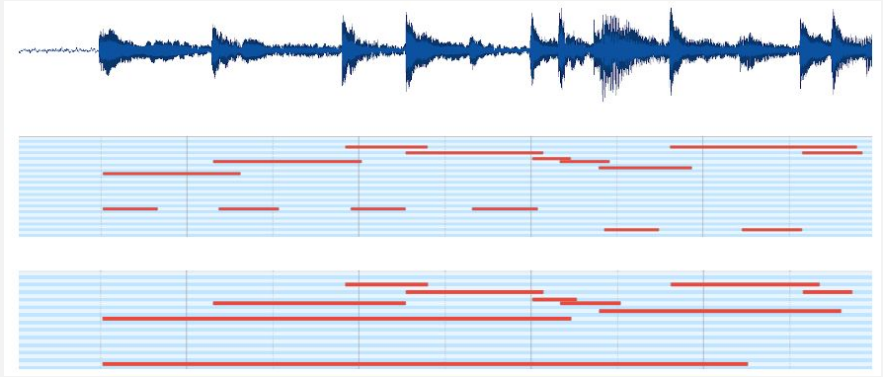
Results

Oregon State University

	Pitch, onset, offset	**Pitch and onset**	Onset only	Offset Only
Precision	0.260	0.646	0.731	0.545
Recall	0.253	0.622	0.700	0.522
F-measure	0.255	0.630	0.710	0.530
Average overlap ratio	0.968	0.677	-	-

	Pitch, onset, and offset			Pitch and onset		
	Precision	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
Hawthorne et al., 2021 [15]	-	-	83.94	-	-	96.01
Hawthorne et al., 2018 [7]	51.32	49.31	50.22	84.24	80.76	82.29
Kelz et al., 2019 [6]	61.93	51.66	56.08	90.15	74.78	81.38
This Work, 2022 →	25.97	25.33	25.50	64.62	62.81	62.97
Kelz et al., 2016 [16]	20.13	27.80	23.14	44.27	61.29	50.94
Sigtia et al., 2016 [13]	17.64	19.71	18.38	44.97	49.55	46.58

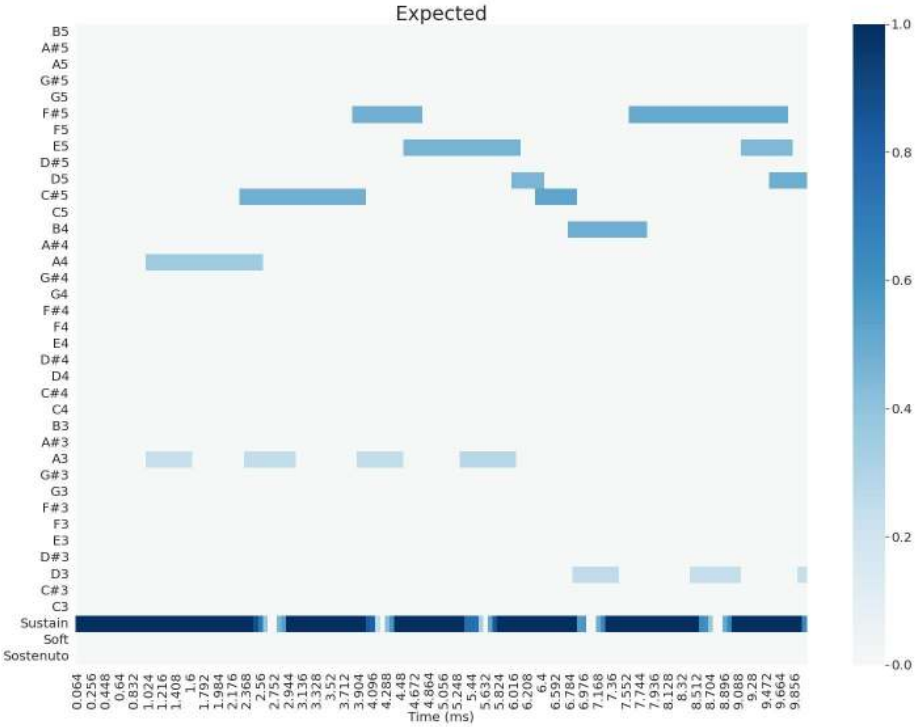
Sonata in A Major, K. 208

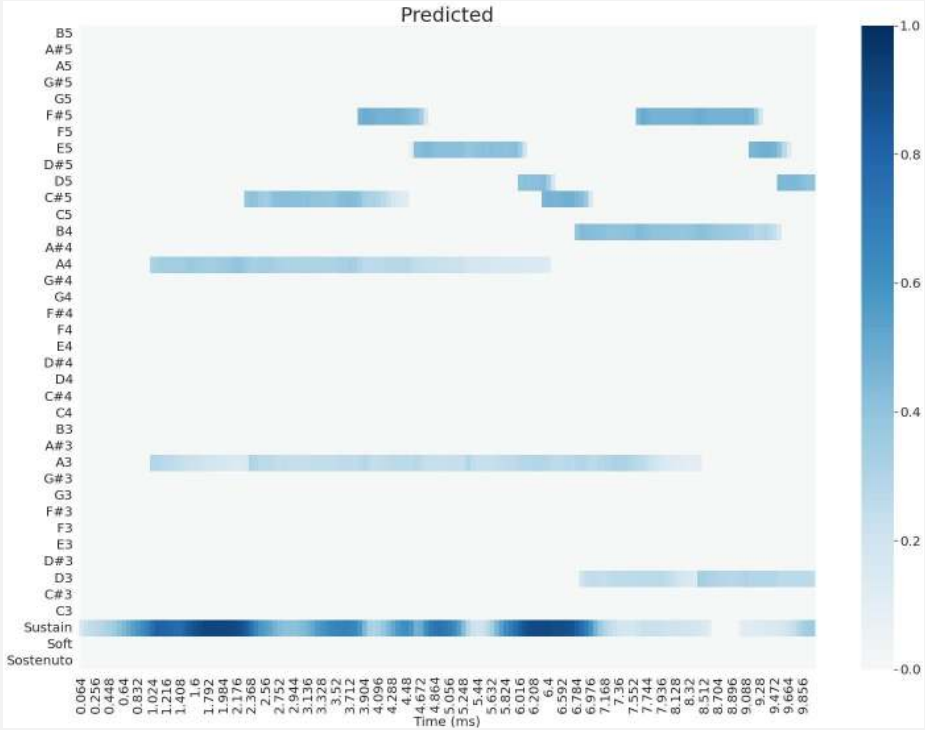


	Pitch, onset, offset	**Pitch and onset**	Onset only	Offset Only
Precision	0.347	0.669	0.823	0.734
Recall	0.267	0.516	0.634	0.556
F-measure	0.302	0.582	0.716	0.639
Average overlap ratio	0.993	0.760	-	-

Sonata in A Major, K. 208

Oregon State University

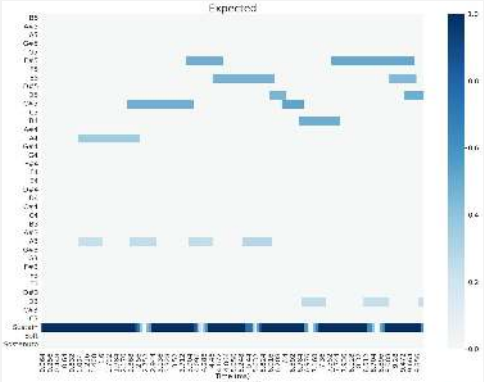




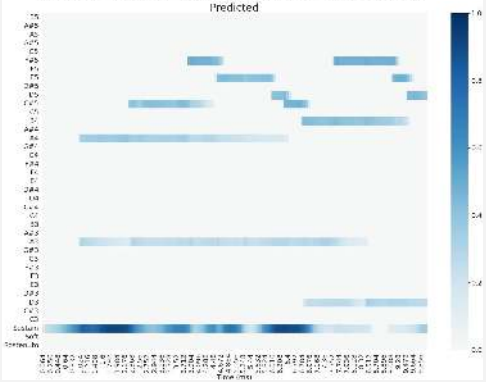
Sonata in A Major, K. 208

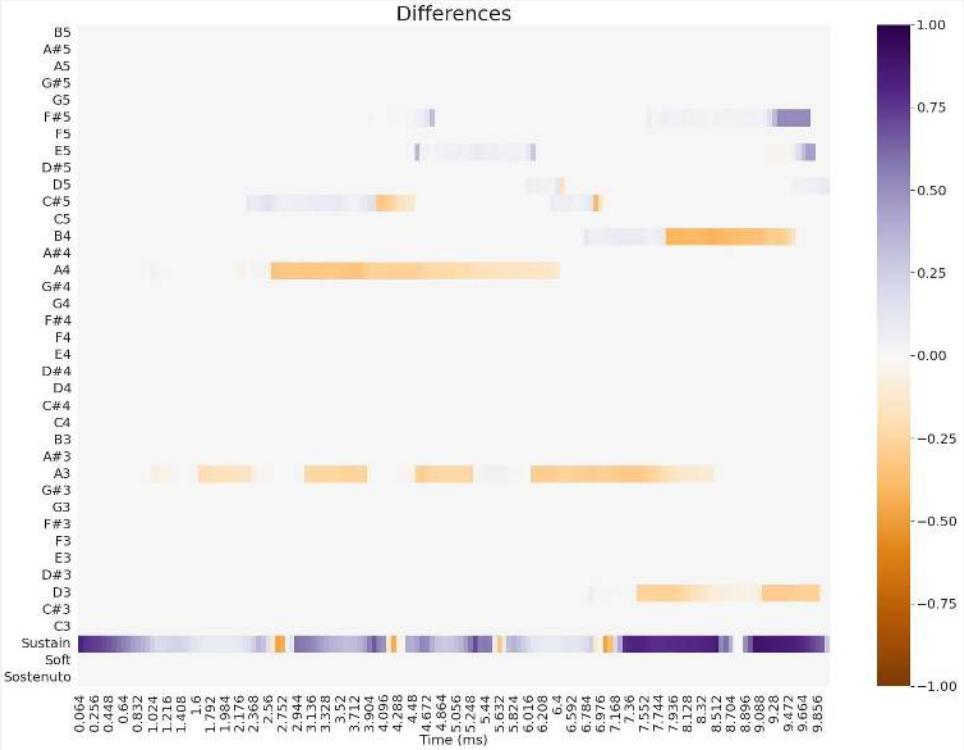
Oregon State University

Expected

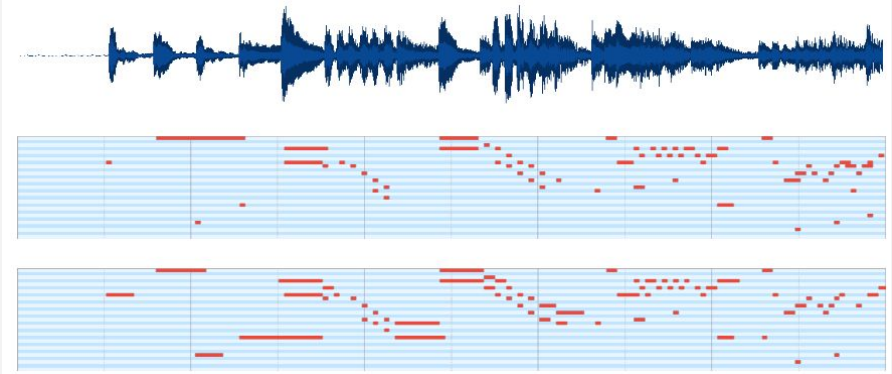


Predicted





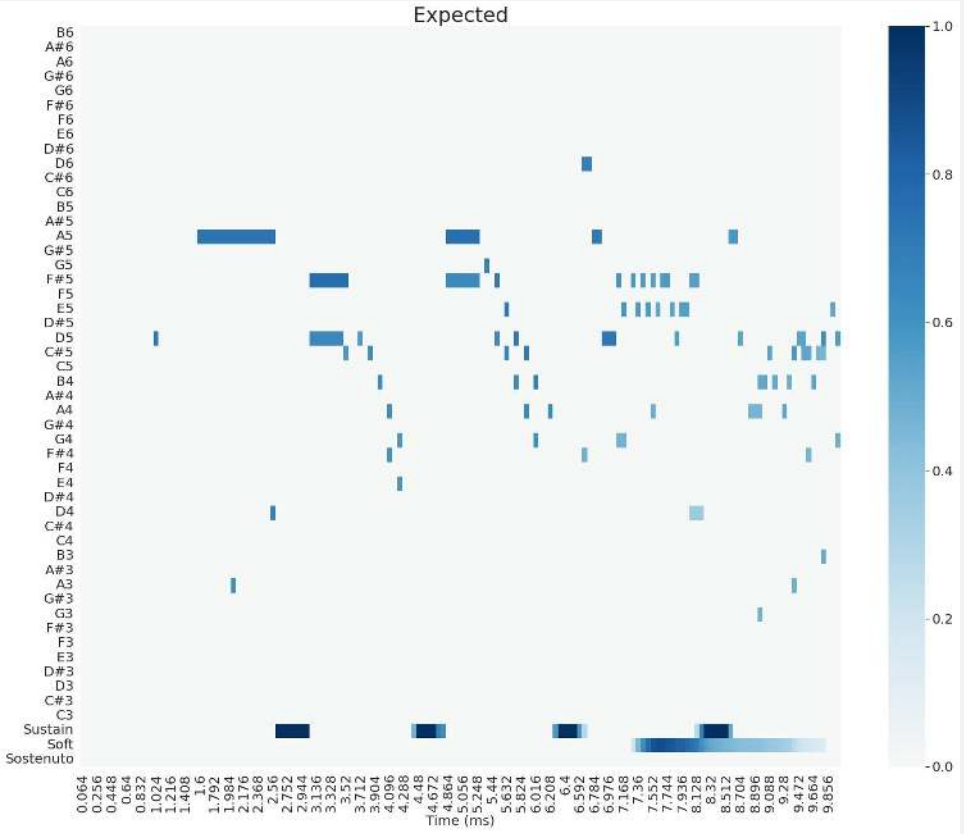
Sonata in D Major, K. 118

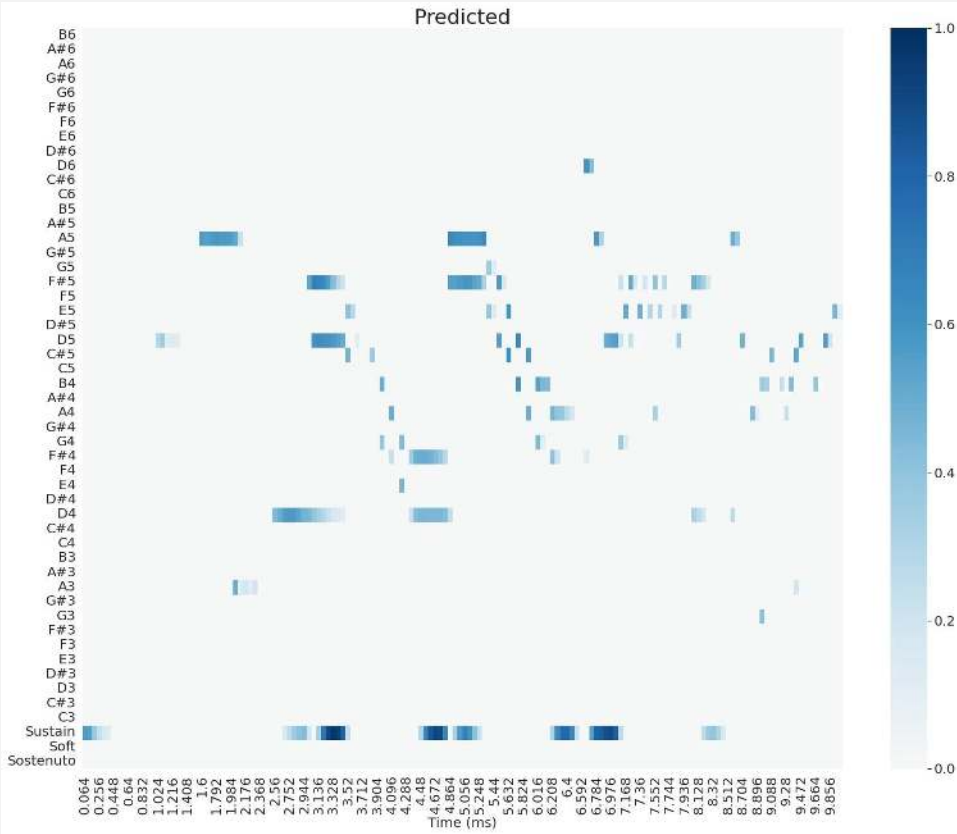


	Pitch, onset, offset	**Pitch and onset**	Onset only	Offset Only
Precision	0.624	0.894	0.894	0.765
Recall	0.589	0.844	0.844	0.722
F-measure	0.606	0.869	0.869	0.743
Average overlap ratio	0.989	0.838	-	-

Sonata in D Major, K. 118

Oregon State University

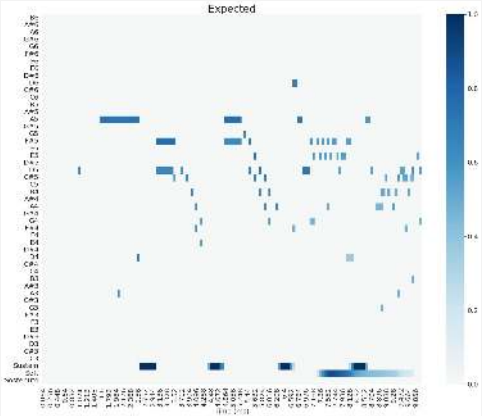




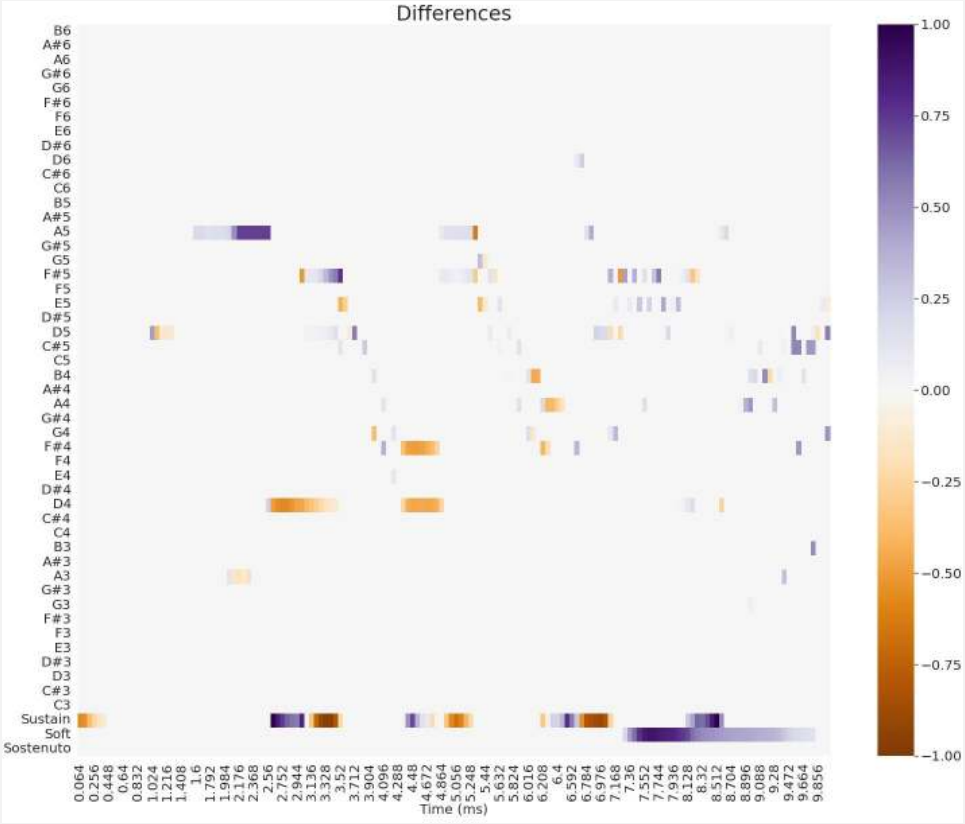
Sonata in D Major, K. 118

Oregon State University

Expected



Predicted



Oregon State University

CONCLUSION

Future Work



- Pedal
  - we included the pedal as features
  - others integrate into training data
- Velocity
  - “velocity smoothing” for playback
  - intelligent or dynamic thresholds
- Evaluation
  - manual evaluation of mistakes
  - threshold for onset/offset

Summary

Oregon State University

- Promising results
  - competitive with older (~5 years) models
  - much simpler pipeline than current state-of-the-art
  - room for refinement and improvement
- Novelty
  - raw audio: no signal processing or feature extraction
  - potential to scale to real-time use

References

[1] M. Marolt, “A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music,” 2004.

[2] J. Nam, J. Ngiam, H. Lee, and M. Slaney, “A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations,” p. 7.

[3] S. Sigtia, E. Benetos, and S. Dixon, “An End-to-End Neural Network for Polyphonic Music Transcription,” p. 14.

[4] S. Sigtia, E. Benetos, et al., “A Hybrid Recurrent Neural Network For Music Transcription,” 2014.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” 2014.

[6] R. Kelz, S. Böck, and G. Widmer, “Deep Polyphonic ADSR Piano Note Transcription,” 2019.

[7] C. Hawthorne et al., “Onsets and Frames: Dual-Objective Piano Transcription,” 2018.

[8] C. Hawthorne et al., “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset,” 2020.

[9] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” 2014.

[10] J. J. Valero-Mas, E. Benetos, and J. M. Iñesta, “Assessing the Relevance of Onset Information for Note Tracking in Piano Music Transcription,” 2017.

[11] S. Bock and M. Schedl, “Polyphonic piano note transcription with recurrent neural networks,” 2012.

[12] A. Cogliati, Z. Duan, and B. Wohlberg, “Piano Transcription With Convolutional Sparse Lateral Inhibition,” 2017.

[13] A. Cogliati, Z. Duan, and B. Wohlberg, “Context-Dependent Piano Music Transcription With Convolutional Sparse Coding,” 2016.

[14] A. Grossman, “Automatic Music Transcription: Generating MIDI From Audio,” p. 6.

[15] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, “Sequence-To-Sequence Piano Transcription With Transformers,” 2021.

[16] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, “On the Potential of Simple Framewise Approaches to Piano Transcription,” 2016.

[17] M. Müller, “Music Representations,” 2015.