Automatic Piano Transcription with Hierarchical Frequency-Time Transformer

Keisuke Toyama[1], Taketo Akama[2], Yukara Ikemiya[3], Yuhta Takida[3], Wei-Hsiang Liao[3], Yuki Mitsufuji[1,3]
[1]Sony Group Corporation, [2]Sony Computer Science Laboratories, [3]Sony AI

SONY

Introduction

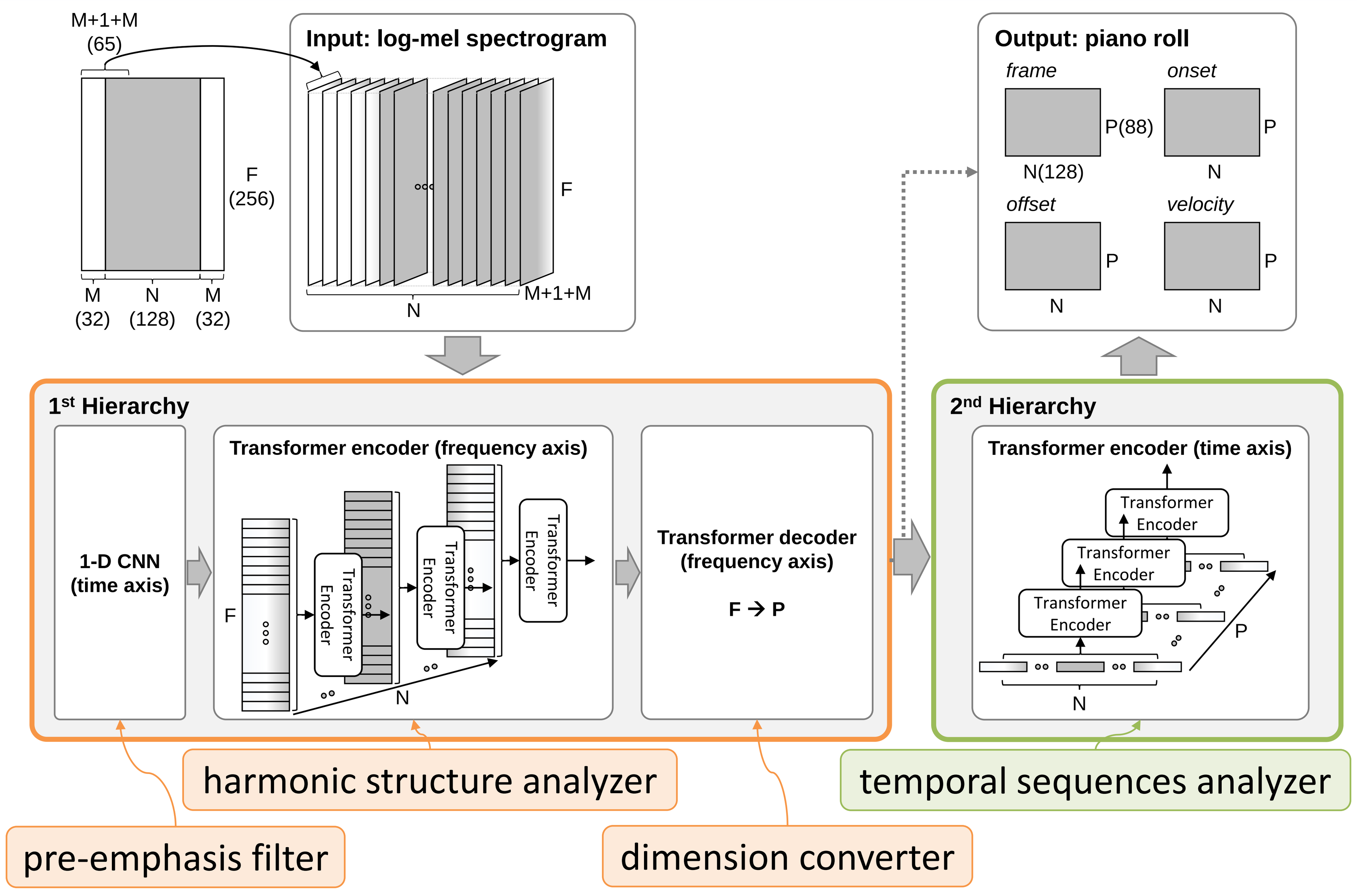
For automatic music transcription (AMT), it is important to analyze
- several harmonic structures that spread in a wide range of frequencies
- temporal sequences of acoustic features in the time axis

→ **self-attention mechanism is a powerful tool to capture the long-term dependency**

We propose an AMT method that uses a 2-level hierarchical frequency-time Transformer architecture
- 1st hierarchy: 1-D CNN (time), 1st Transformer encoder (frequency), Transformer decoder (frequency)
- 2nd hierarchy: 2nd Transformer encoder (time)

hFT-Transformer

Methods

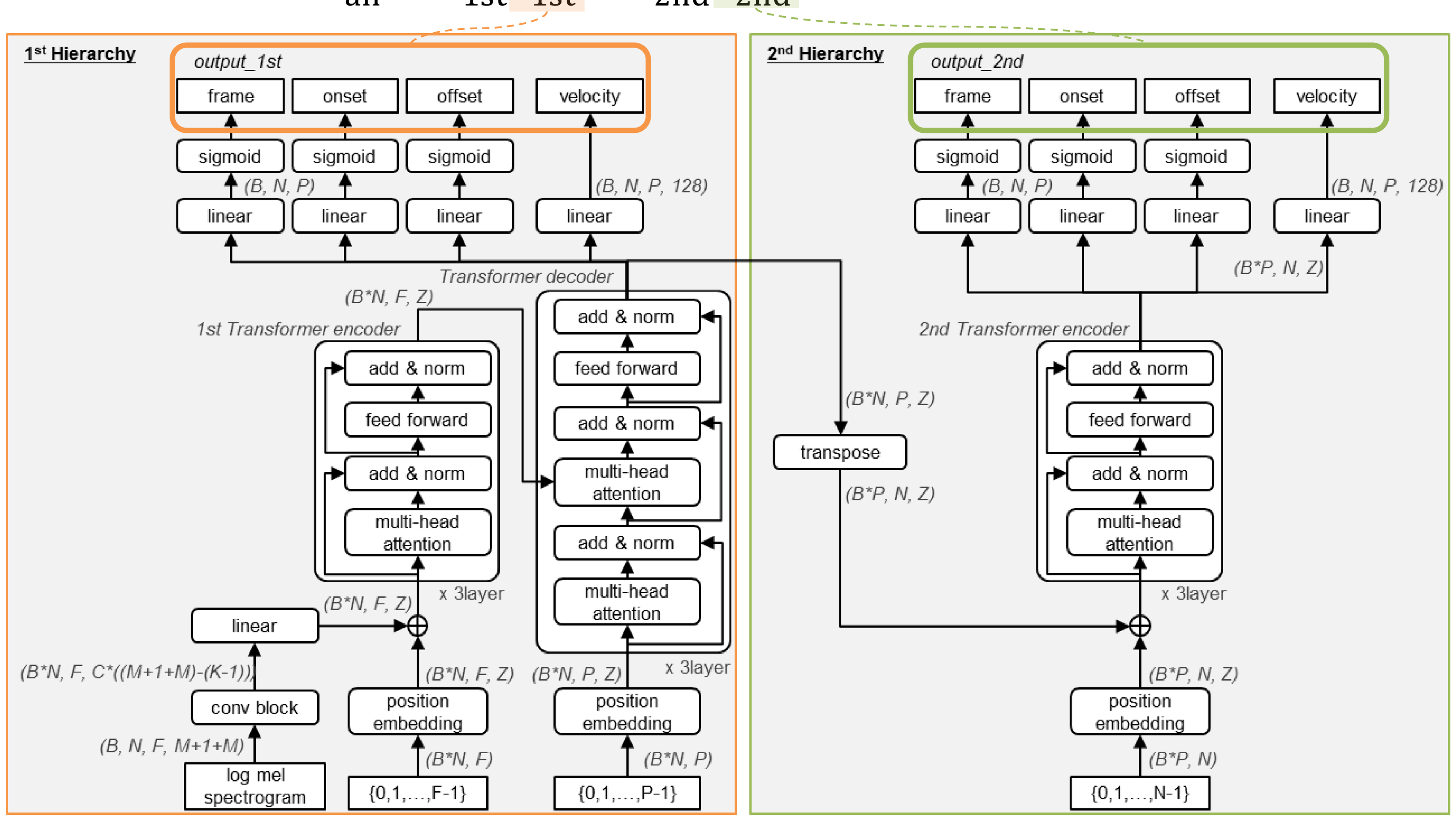


Loss function

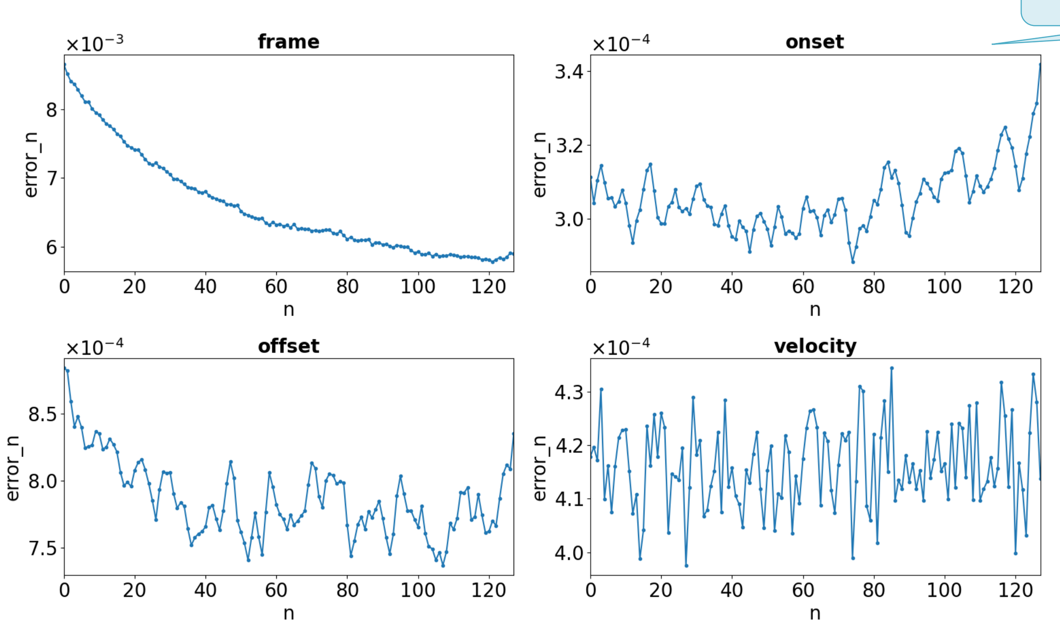
- frame, onset, offset: binary cross-entropy
- velocity: 128-category cross-entropy

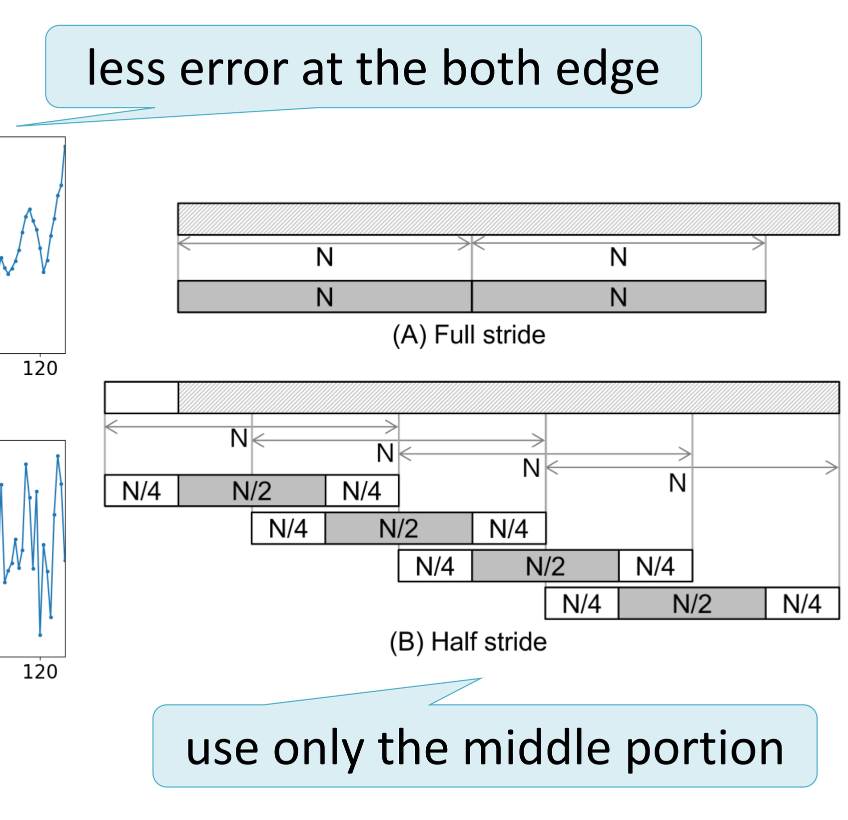
$$L = L_{\mathrm{bce}}^{\mathrm{frame}} + L_{\mathrm{bce}}^{\mathrm{onset}} + L_{\mathrm{bce}}^{\mathrm{offset}} + L_{\mathrm{cce}}^{\mathrm{velocity}}$$

$$L_{\mathrm{all}} = \alpha_{\mathrm{1st}} L_{\mathrm{1st}} + \alpha_{\mathrm{2nd}} L_{\mathrm{2nd}}$$



Half-stride strategy





$$error_n^{<m>} = \frac{1}{IP} \sum_{i=0}^{I-1} \sum_{p=0}^{P-1} \left(y_{i,n,p}^{<m>} - \hat{y}_{i,n,p}^{<m>}\right)^2$$

Experimental Results

Dataset (Piano)

- MAPS (train/valid/test=8.3/4.4/5.5 hours)
- MAESTRO v3.0.0 (159.2/19.4/20.0 hours)

Results

- outperformed the other existing methods
- half-stride strategy is effective

Dataset	Method	Half-stride	Params	Frame	Note	Note Offset	Note Offset&Velocity
MAPS	Onsets&Frames [7]		26M	78.30	82.29	50.22	35.59
	ADSR [10]		0.3M	77.16	81.38	56.08	-
	hFT-Transformer		5.5M	<u>82.67</u>	<u>85.07</u>	<u>66.03</u>	<u>47.92</u>
	hFT-Transformer	✓	5.5M	**82.89**	**85.14**	**66.34**	**48.20**
MAESTRO v3.0.0	Seq2Seq [3]		54M	-	96.01	83.94	82.75
	HPT-T [2]		-	90.09	96.77	83.20	81.90
	Semi-CRFs [12]		9M	90.75	96.11	88.42	87.44
	HPPNet-sp [5]		1.2M	<u>93.15</u>	97.18	83.80	82.24
	hFT-Transformer		5.5M	93.02	<u>97.43</u>	<u>90.32</u>	<u>89.25</u>
	hFT-Transformer	✓	5.5M	**93.24**	**97.44**	**90.53**	**89.48**
	SpecTNT (*)		-	-	(96.9)	-	-
	PerceiverTF (*)		-	-	(96.7)	-	-

Evaluation results on MAPS/MAESTRO test dataset (**bold**: best score, <u>underline</u>: 2nd best score)
(*: reported in "Multitrack Music Transcription with a Time-Frequency Perceiver," in ICASSP2023)

Ablation Study

Model	1st Hierarchy: CNN	1st Hierarchy: 1st Encoder	1st Hierarchy: Converter	2nd Hierarchy: 2nd Encoder	Output	Params	Frame	Note	Note Offset	Note Offset& Velocity
hFT-Transformer	1-D	✓	Decoder	✓	2nd	5.5M	**91.09**	**96.72**	**84.42**	**75.95**
1-F-D-N	1-D	✓	Decoder	n/a	1st	3.9M	90.09	<u>95.95</u>	80.23	<u>71.78</u>
2-F-D-T	2-D	✓	Decoder	✓	2nd	6.1M	67.52	31.10	20.88	13.50
1-F-L-T	1-D	✓	Linear	✓	2nd	3.4M	<u>90.99</u>	95.79	<u>82.98</u>	69.34

Evaluation results of ablation study on MAPS validation dataset

2nd Transformer encoder in time axis (vs 1-F-D-N)

- presumably helpful in offset estimation

Complexity of convolutional block (vs 2-F-D-T)

- 2-D convolution block may over aggregate the spectral information

Converter (vs 1-F-L-T)

- effective in velocity estimation

Coefficients of loss functions

$(\alpha_{\mathrm{1st}}, \alpha_{\mathrm{2nd}})$

(1.0,1.0) pair yields the best score